Design and Analysis of Sequential Multiple Assignment Randomized Trial for Comparing Multiple Adaptive Interventions

Xiaobo Zhong

Submitted in partial fulfillment of the
requirements for the degree
of Doctor of Public Health
in the Mailman School of Public Health

Columbia University

2018

ABSTRACT

Design and Analysis of Sequential Multiple Assignment Randomized Trial for Comparing Multiple Adaptive Interventions

Xiaobo Zhong

The research of my dissertation studies the methods of designing and analyzing sequential multiple assignment randomized trial (SMART) for comparing multiple adaptive interventions. As a SMART typically consists of numerous adaptive interventions, inferential procedures based on pairwise comparisons of all interventions may suffer substantial loss in power after accounting for multiplicity. I address this problem using two approaches. First, I propose a likelihood-based Wald test, study the asymptotic distribution of its test statistics, and apply it as a gate-keeping test for making an adaptive intervention selection. Second, I consider a multiple comparison with the best approach by constructing simultaneous confidence intervals that compare the interventions of interest with the truly best intervention, which is assumed to be unknown in inference; an adaptive intervention with the proposed interval excluding zero will be declared as inferior to the truly best with a pre-specified confidence level. Simulation studies show that both methods outperform the corresponding multiple comparison procedures based on Bonferroni's correction in terms of the power of test and the average width of confidence intervals for estimation. Simulations also suggest desirable properties of the proposed methods. I apply these methods to analyze two real data sets. As part of the dissertation, I also develop a user-friendly R software package that covers many statistical work related to SMART, including study design, data analysis and visualization. Both proposed methods can be implemented by using this R package. In the end of the dissertation, I show an application of designing a SMART to compare multiple patient care strategies for depression management based on one of the proposed methods.

Contents

Acknowledgements

The path can not have been completed without the amazing synchronicities that came together to bring me to this point. Over the past five years and all the hard work of my study, I have been most fortunate to meet such amazing and dedicated scholars and staffs at Columbia University. First and foremost, I would like to thank my advisor, Professor Ying Kuen Cheung, for his kindness, help and encouragement. His erudition, diligence and thoroughness as a outstanding scholar inspired me to pursue statistical methodological research on clinical trial and brought me to a new journey of my life. I am grateful for all my dissertation committee members, including Professors Bin Cheng, Min Qian, Ian Kronish and Professor Bruce Levin, who is the committee chair and the past chair of department of biostatistics at Columbia University, for their help during my dissertation research. During my years as a doctoral student in Columbia University, it has been my honor to attend courses taught by Professors Ying Wei, Shing Lee, Shuang Wang, Yuanjia Wang, Arindam RoyChoudhury, Xinhua Liu, Prakash Gorroochurnm, John Thompson, Wei-Yann Tsai, Srikesh Arunajadai, Jeff Goldsmith and Cheng-Shiun Leu. The data used in my dissertation were collected by the Center for Behavioral Cardiovascular Health at Columbia University Medical Center and the Center for Health Communications Research at University of Michigan. The funding to my research is from the Merit Doctoral Scholar Award at Columbia University and the NIH grant R01MH109496: Novel Methods for Evaluation and Implementation of Behavioral Intervention Technologies for Depression, of which Professor Ying Kuen Cheung is the principle investigator. Finally, the completion of this dissertation would not have been possible without the support of Dr. Parameswaran Hari, who encouraged and supported me to joined my doctoral program, and without the supports of my mother Qinfen Zeng, father Xiankun Zhong, daughter Marta Zhong and of my wife Qixuan Chen.

Chapter 1 Adaptive Intervention and SMART Design

My dissertation proposes methods for designing and analyzing randomized clinical trial, specifically, sequential multiple assignment randomized trial (SMART), with the objective to compare multiple adaptive interventions. I introduce in this chapter the background and the motivation of my research. First, I introduce the concepts of adaptive intervention in Chapter 1.1 and SMART in Chapter 1.2. And then I briefly compare SMART with some other clinical trial designs that share some common features with SMART in Chapter 1.3 . This chapter ends with the motivation of my dissertation research in Chapter 1.4.

1.1 Adaptive Intervention

Personalized medicine, or called precision medicine, has become increasingly interesting in statistical research in the past 2 decades. Although the concept of personalized medicine varies by literatures and research communities, the basic idea is somehow involving individual-level information in treatment selection (Kosorok, 2016). An adaptive intervention (AI) is a multi-stage treatment strategy consisting of a sequence of treatment selections, one per stage of treatment for a patient, which can be repeatedly adjusted according to the individual's ongoing clinical information, such as the treatment history and the responses to the previous treatments. AIs have been widely used in managing chronic diseases, such as cancer and depression, with the belief that the long-term response of an individual can be optimized by adjusting the treatment selection as a function of time-varying personalized data. Thus, it fits in the larger paradigm of personalized medicine (Chakraborty and Murphy, 2014). There are multiple terms that have been used in liter-

atures to describe such a multi-stage treatment strategy, including dynamics treatment regime (Robins, 1986), adaptive treatment strategy (Lavori and Dawson, 2007; Murphy, 2005), treatment policy (Lanceford, Davidian and Tsiatis, 2012; Wahed and Tsiatis, 2004), individualized treatment rule (Van der Lann and Peterson 2007), and adaptive intervention (Collins, 2014). I use *adaptive intervention* in my dissertation.

Figure 1.1 gives an example of two-stage AI used for untreated diffuse large B-Cell lymphoma (DLBCL) patients (Habermann et al., 2006). The goal is to help untreated DLBCL patients to achieve durable complete remission. Under this AI, patients were first given cyclophosphamide, doxorubicin, vincristine, and prednisone (CHOP) at baseline for 4 weeks as stage-1 treatment. At the end of Stage 1, an intermediate evaluation was given to each patient and they were classified as respondents and non-respondents based on the results of assessments. A patient who achieved complete remission (CR) or partial remission (PR) within 4 weeks was classified as respondent and given Rituximab for maintenance at Stage 2. On the other hand, a patient with stable disease, progression or relapse was classified as non-respondent and given granulocyte colony-stimulating factor (G-CSF) as supportive care at Stage 2. Comparing to a non-adaptive strategy that simply gave patients CHOP followed by Rituximab or CHOP followed G-CSF, the treatment selection at Stage 2 under this AI took into account the individual-level information using the response to CHOP as surrogate, thus could potentially improve the long-term health outcome of an individual. A patient under this AI could possibly follow one of two treatment sequences indicated by the treatments selected and the response to stage-1 treatment. One sequence was "CHOP $\rightarrow$ response $\rightarrow$ Rituximab" and the other is "CHOP $\rightarrow$ no response $\rightarrow$ G-CSF". Which sequence was actually followed by a patient depended on the value of intermediate outcome observed in clinical practice. In other

words, the treatment sequence received by a patient given an AI is completely specified by the observed intermediate response.

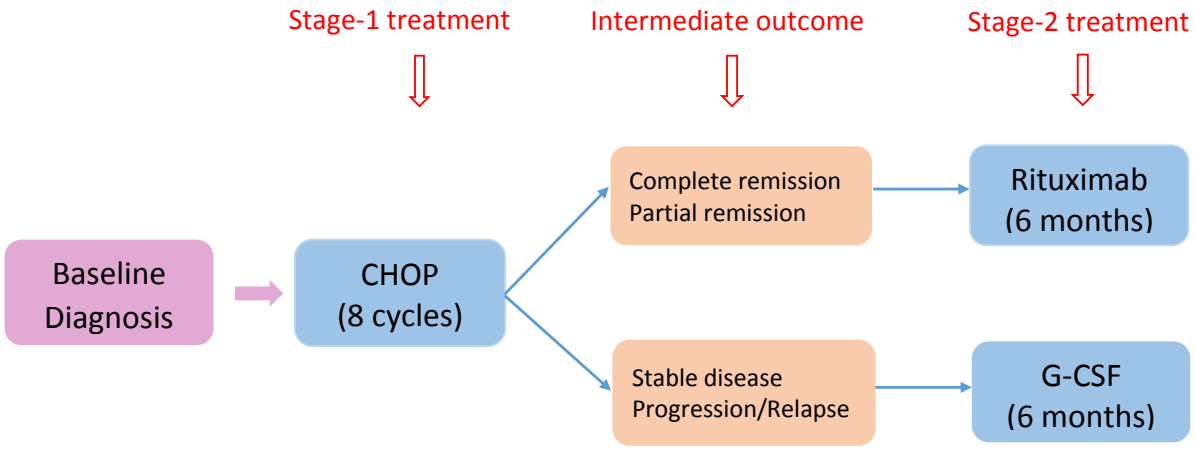


Figure 1.1 An adaptive intervention for patients with diffuse large B-cell lymphoma.

1.2 Sequential Multiple Assignment Randomized Trial

Sequential Multiple Assignment Randomized Trial (SMART) is the randomized clinical trial (RCT) design that randomly assigns a collection of AIs to patients. Figure 1.2 is an example of SMART design with the objective to compare 4 two-stage AIs for untreated DLBCL patients (Habermann et al., 2006). Although CHOP has been considered as the standard treatment for the front-line chemotherapy for untreated DLBCL patients, a clinical investigation found that CHOP plus rituximab, a chimeric antibody that targeted CD20 B cells, could potentially improve the overall response of the front-line therapy in this population (Czuczman, Grillo-Lopez and White, 1999). Meanwhile, for those DLBCL patients who successfully achieved CR/PR in the front-line therapy, both rituximab and observation are commonly used for maintenance in practice. In this trial, patients

were first randomized to receive either CHOP or rituximab plus CHOP (R-CHOP) as induction for 6-8 cycles at Stage 1. CR/PR patients who completed induction therapy were randomly assigned to receive maintenance rituximab or observation for 6 months at Stage 2, while those who failed to achieved CR/PR were given G-CSF as supportive care according to the guideline of American Society of Clinical Oncology (American Society of Clinical Oncology, 1996). A patient who completed this trial could followed one out of 6 possible treatment sequences, including (1) "CHOP $\rightarrow$ Response $\rightarrow$ Rituximab", (2) "CHOP $\rightarrow$ Response $\rightarrow$ Observation", (3) CHOP $\rightarrow$ No response $\rightarrow$ G-CSF, (4) "R-CHOP $\rightarrow$ Response $\rightarrow$ Rituximab", (5) R-CHOP $\rightarrow$ Response $\rightarrow$ Observation, and (6) R-CHOP $\rightarrow$ No response $\rightarrow$ G-CSF", depending on the results of sequential randomizations and the value of intermediate outcome observed on the patient. Consequently, it provided data that allowed to compare 4 AIs. By virtue of sequential randomization, the assumption of ignorable treatment holds in this case and thus the conclusion can be referred as a RCT-based evidence for practice guideline.

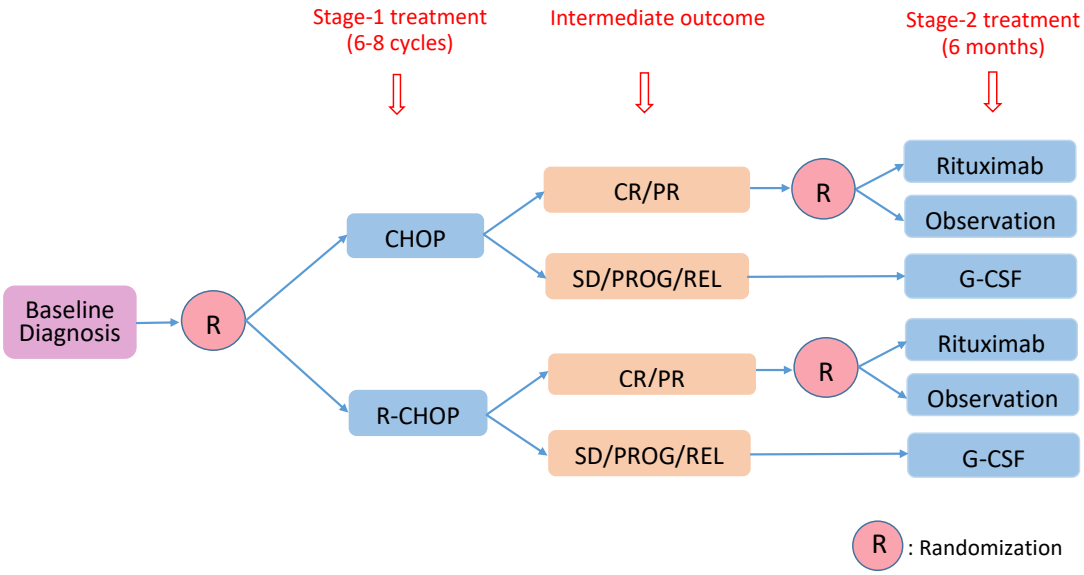


Figure 1.2. Two-stage SMART design for untreated DLBCL patients. CR: complete remission; PR: partial remission; SD: stable disease; PROG: progression; REL: Relapse; CHOP: cyclophosphamide, doxorubicin, vincristine, and prednisone; R-CHOP: rituximab plus CHOP ; G-CSF: granulocyte colony-stimulating factor.

SMART is an efficient design for early phase clinical trials when we are interested in evaluating a collection of AIs. A common goal of such trials is to identify one promising or several near-best candidates of AIs and move forward to a confirmatory study. Suppose we are interested in evaluating the effects of 4 AIs that were defined in the DLBCL trial (cf. Figure 1.2) and use a traditional RCT design that simply assigns patients randomly to 4 independent groups at baseline, each of which corresponds to one AI of interest. To obtain n patients for each AI, assuming equal randomization probabilities, we need $4n$ patients in total. However, by using a SMART design as shown in Figure 1.2, given the same total sample size of $4n$, we can have $(1+p)n$ patients for each AI, where p is the response rate of stage-1 treatment. Another attractive feature of SMART is that the sequential randomization provides data that allow to study multiple research questions.

Because there are multiple randomization points, we can actually viewed SMART as a combination of several small RCTs given different conditions, and thus use the data to explore multiple secondary questions. For example, the design in Figure 1.2 provides data that allow no only to compare 4 AIs, but also to make 3 other comparisons (1) CHOP vs. R-CHOP at baseline, (2) maintenance rituximab vs. observation for patients who responded to CHOP, and (3) maintenance rituximab vs. observation for patients who responded to R-CHOP. The reason I called these "secondary" research questions is that I assume the sample size calculation of this SMART is based on the primary question of comparing 4 AIs. Thus, it may not have enough sample sizes to achieve certain powers under the strictly controls of the targeted error rates for other questions. But the design itself still provides valuable information for these questions. My dissertation research focuses on the methods for comparing multiple AIs embedded in SMART, which are applied to early phase trials with the objective to collect information leading to a final confirmatory study for AI research.

1.3 SMART and Other Designs

In this chapter, I compare SMART and several RCT designs that share some common features with SMART more or less. The similarities between SMART and these designs sometimes cause confusions for clinical trialists who get to know SMART at the beginning. By making these comparisons, I further demonstrate the unique features of SMART in the family of RCT designs.

Crossover design is a repeated measurement RCT design such that individuals receive varying treatment sequences across multiple stages (Brown, 1980). Operationally, the

sequential feature makes SMART somewhat similar to the crossover design. However, the motivation of SMART differs from that of the crossover design completely. A crossover design is motivated to improve the efficiency and reduce the sample size of a study aiming to compare non-adaptive interventions. But a SMART aims to compare adaptive interventions. Also, the treatment assignment in a crossover study only depends on the results of randomization at baseline. For example, in a crossover study comparing the treatment effects of A versus B, every patient is assigned to receive either a sequence of "A $\rightarrow$ B" or "B $\rightarrow$ A" at baseline. Once the treatment program initiates, every patient follows the assigned treatment sequence to the end of the study unless he or she drops out. In SMART, only the set of decision rules are pre-specified at baseline. It is unlikely to know the treatment sequence actually received by a patient until the last decision is made. The major pitfall of a crossover design is *carryover effect* such that the prior treatment effect A (or B) may be confounded with the succeeding treatment effect B (or A). To avoid the contamination of carryover effects, a crossover design typically inserts a unique component called *washout period* between two treatment periods within a subject. The key of setting up a washout period is to control the time long enough to diminish any possible carryover effect. Nevertheless, a SMART aims to assess the effects of AIs, which could possibly be a result of delay effects of early treatments. Therefore, synergistic interactions between multi-stage treatments are of interest and thus a SMART design typically does not consider to rule out carryover effect.

Adaptive design uses interim data of a study to modify some design aspects (e.g., randomization scheme) of a trial as it continues, without undermining the validity and integrity of the study (Gallo et al., 2006). Comparing to SMART, adaptive design is a broader concept that covers a family of RCT designs paralleled to SMART, includ-

ing *continual reassessment design* (Garrett-Mayer, 2006), *Seamless design* (Maca et al., 2006), *enrichment design* (Wang Hung and O'Neill, 2009), *sample size recalculation design* (Proschan, 2009) and *adaptive randomization design* (Zhang and Rosenberger, 2012), all sharing a common feature of having opportunities to modify one or several design aspects based on the results of interim analysis (Coffey et al., 2012). Adaptive design is motivated by improving the overall quality of care for trial participants in favor of the treatments showing better efficacy or less toxicity during the early period of a trial and thus the between-subject information is adapted. On the other hand, the treatment assignment of a patient in SMART is adapted to the within-subject information, while the common design elements (e.g., sample size, randomization scheme, etc.) are pre-specified to the trial and will not be modified during the study. Cheung et al. (2015) proposed a design, *Sequential Multiple Assignment Randomization Trial with Adaptive Randomization (SMART-AR)*, which can improves the quality of patient care by adapting some design parameters in a SMART framework. In SMART-AR, subjects received treatments based on sequential randomization as in a classic SMART and the entire study was designed into multiple stages so that some design parameters (e.g. randomization probabilities) can be modified based on the interim analysis in the same clinical trial. The simulation results indicated that overall quality of patient care can be improved by adjusting the design parameters of adaptive randomization and the elites of SMART remain in the trial with SMART-AR design. Further exploration of SMART-AR is open to researchers.

Factorial design decomposes the variation of primary outcome into the main effects and possible interaction effects of one or several factors. Factorial design is a classic experimental design that has been widely used not only in clinical trial research, but also in a variety of areas, such as agriculture (Fisher, 1926), engineering (Giachetti et

al., 2013), and marketing (Holland and Cravens, 1973). The experience about factorial design can help to understand SMART design better. We can view SMART as a factorial design under which time and treatment decisions play the essential roles (Chakraborty et al., 2009). For example, in the DLBCL trial mentioned in Chapter 1.2, we can explain the variation of primary outcome by 3 factors (stage-1 treatment, intermediate response and stage-2 treatment) so as to make it a typical 3-way factorial design.

1.4 Motivation of Research

My dissertation research is motivated by developing statistical methods for SMART with the goal of comparing multiple AIs. SMART provides a very flexible framework that allows varying design structures according to the AIs of interest in study, featured by the numbers of stages, treatment options and intermediate response categories. Figure 1.3 shows a modified design of the DLBCL trial in Figure 1.2, in which the intermediate outcomes are separated into three categories and only PR patients undergo the second randomization. Such a design leads to data that allow to compare 4 AIs, defined as (1) CHOP followed by observation (for CR patients) or rituxamab (for PR patients) or G-CSF (otherwise), (2) CHOP followed by observation (CR/PR) or G-CSF (otherwise), (3) R-CHOP followed by observation (CR) or rituxamab (PR) or G-CSF (otherwise) and (4) R-CHOP followed by observation (CR/PR) or G-CSF (otherwise). While Figure 1.4 gives another modified SMART design of the DLBCL trial that leads to data that allow to compare up to 8 two-stage AIs similar to those defined in the DLBCL trial.

Figure 1.3. Modified SMART design (A) of the DLBCL trial. CR: complete remission; PR: partial remission; SD: stable disease; PROG: progression; REL: Relapse; CHOP: cyclophosphamide, doxorubicin, vincristine, and prednisone; R-CHOP: rituximab plus CHOP ; G-CSF: granulocyte colony-stimulating factor.

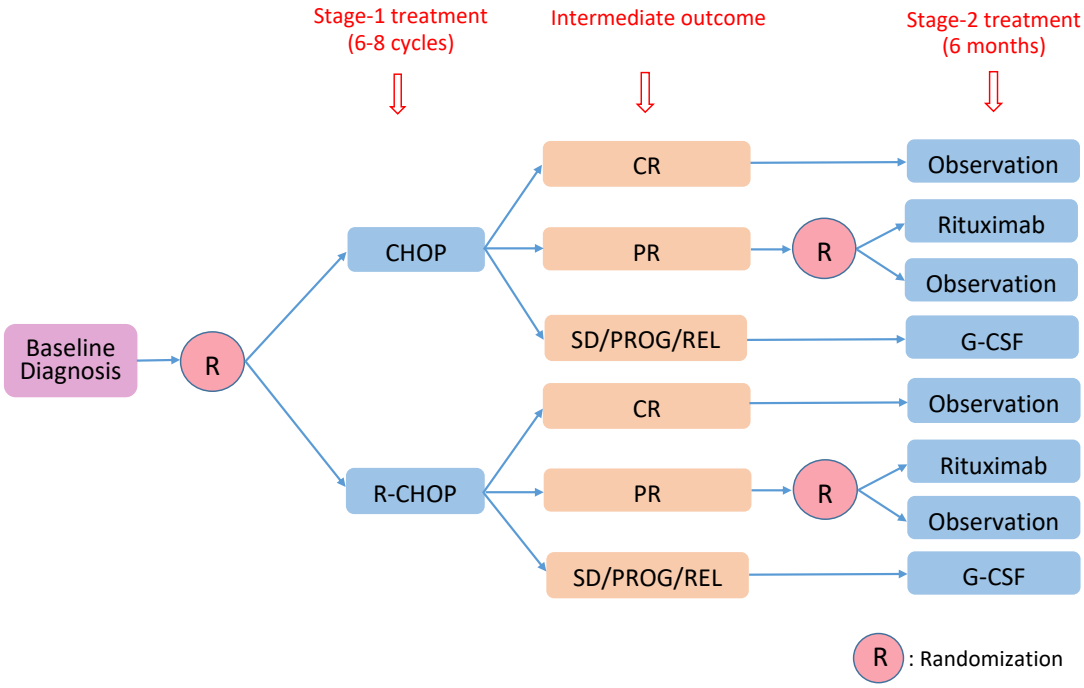
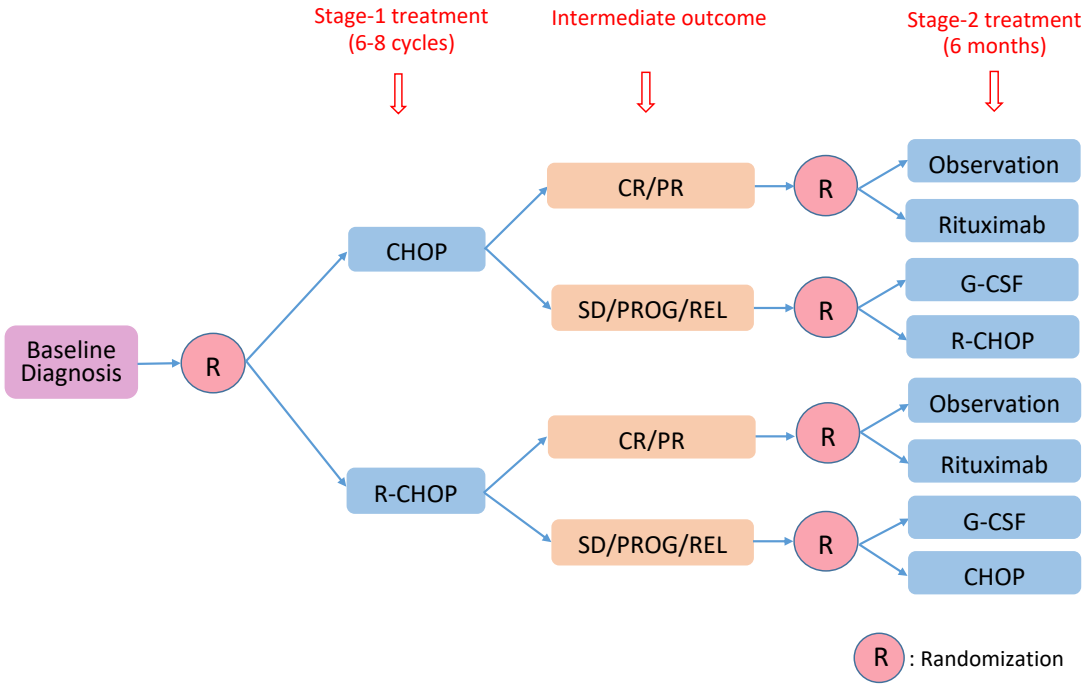


Figure 1.4. Modified SMART design (B) of the DLBCL trial. CR: complete remission; PR: partial remission; SD: stable disease; PROG: progression; REL: Relapse; CHOP: cyclophosphamide, doxorubicin, vincristine, and prednisone; R-CHOP: rituximab plus CHOP ; G-CSF: granulocyte colony-stimulating factor.



It is interesting that the total number of AIs embedded in a SMART could increases exponentially as the design structure gets more complicated, which typically reflects more treatment stages, more treatment options or more intermediate response categories. For example, when we compare the designs between Figure 1.2 versus Figure 1.4, as the number of treatment options for patients who fail in stage-1 treatment increases from one to two, the total number of AIs increases from 4 to 8. The sequential randomization automatically generates SMART data of more than one pair of AIs. Unlike in a simple RCT design aiming to compare non-adaptive treatments such that all the intervention groups are independent, some AIs in a SMART are partially overlapped by certain treatment sequences. For example, AI (1) and AI (2) described in the previous paragraph are overlapped by two treatment sequences, which are "CHOP $\rightarrow$ CR $\rightarrow$ observation" and "CHOP $\rightarrow$ stable disease/progression/relapse $\rightarrow$ G-CSF". As results, the estimated values of some AIs embedded in a SMART are correlated and the correlations vary by design structures. Intuitively, patients who complete the treatment sequence of "CHOP $\rightarrow$ CR $\rightarrow$ G-CSF" in Figure 1.3 contribute data in estimating both AI (1) and AI (2) in the previous paragraph. Therefore, the covariance between the estimated values of these two AIs needs to be taken into consideration in the inference for comparing two AIs. A well-developed method should be applicable to SMART with varying design structures.

The rest of my dissertation is organized as follows. In Chapter 2, I will propose a Wald-type omnibus test, which can be applied as a gate-keeping test to select the best AI embedded in SMART. This method is recommended in a early phase trial with the goal to select one promising for further confirmatory study. I will show an application using a real data set extracted from a smoking cessation trial. In Chapter 3, I will propose a method to construct a set of simultaneous confidence intervals, called *Multiple Comparison with*

the Best Simultaneous Confidence Intervals, which can be used as a filter to identify the inferior AIs and eliminate it from moving forward in a series of experimental studies. I will show an application using the data collected from a depression management trial. I have developed an R package, named *SRT (**S**equential **R**andomized **T**rial)* for implementing SMART design. This package contains most commonly used methods in SMART settings and both the methods proposed in my dissertation have also been built in this package. I will introduce this R package and illustrate the usage of this R package in Chapter 4. In Chapter 5, I will report an exploratory study that illustrates using SMART design to improve the efficiency comparing to a traditional randomized clinical trial in a depression prevention study. This dissertation ends up with conclusion and some discussions about future directions in Chapter 6.

Chapter 2 A Gate-keeping Test for Selecting Adaptive Intervention under General SMART Designs

2.1 Introduction

An adaptive intervention (AI) consists of a sequence of treatment decisions made based on a patient's historical clinical information, such as treatment history and responses to previous treatments. AI has long been a common treatment strategy in the clinical practice for cancer, mental disorders, and many other chronic conditions. Evaluation of AIs in experimental settings, on the other hand, has been considered only recently in the context of sequential multiple assignment randomized trial (SMART; Rush et al., 2004; Thall et al., 2007). A SMART may be viewed as a clinical trial design that provides data about a collection of randomly assigned AIs that may overlap in terms of treatment decisions. In many situations, a SMART could facilitate the selection and the prioritization of interventions in a series of experimental studies that lead to a confirmatory trial (Murphy, 2005). As such, a natural research objective of a SMART is to determine whether or not an AI should be moved forward for further investigation. Specifically, in this chapter, I consider a method related to the selection of AIs embedded in a SMART.

By virtue of randomization upon observing treatment history and tailoring response, the value of each AI can be consistently estimated using G-computation under structural nested models (Robins, 1986; Lavori et al., 2007) and inverse probability weighted estimation under the marginal mean models (Murphy et al., 2001; Orellana, Rotnitzky, and Robins, 2010). Thus, the best AI embedded in a SMART may be selected by comparing the estimated values of all AIs embedded in the SMART. This approach entails multi-

ple pairwise comparisons of AIs. As a main concern in a randomized clinical trial is to protect against false positive finding (Hochberg and Tamhane, 1987), multiplicity adjustments are necessary when numerous AIs are evaluated in a comparative fashion. While Bonferroni method is a versatile approach that can be directly applied to the multiple comparison problem in a SMART, it is also known to be conservative. Furthermore, such a problem is magnified when we compare multiple AIs using SMART data. Because the total number of pairwise comparisons needed to be adjusted in SMART sometimes increases exponentially as the design structure gets more complicated. For example, Figure 1.4 is a modified of the DLBCL trial. Comparing to the original design shown in Figure 1.2, one more option has been added to the stage-2 treatment for patients who failed to achieved PR/CR at Stage 1 in the modified SMART design. Consequently, the total number of AIs increases from 4 to 8 so that the number of pairwise comparisons needed to be adjusted by Bonferroni's method increases from 6 to 28. As an alternative, one could account for multiplicity by using a gate-keeping approach whereby the selection of AI will be made only when the hypothesis of no difference among the AIs of interest is rejected. Several omnibus tests have been proposed in the literatures; see, for example, Orellana et al. (2010), Nahum-Shani et al. (2012), and Ogabagaber, Karp and Wahed (2016). Using a gate-keeping approach, one can justify the total sample size of a SMART formally with respect to the targeted type I error rate and the power of the omnibus test. While most sample size formulae for SMARTs were derived for comparing two AIs with or without overlap in treatment decisions (e.g., Murphy, 2005; Oetting et al., 2009), powering a study based on an omnibus test is arguably more relevant than one based on pairwise comparisons, because the former accounts for all embedded AIs. Indeed, in many cases, it is practically challenging for a trialist to choose only one pair of AIs to

power a SMART due to lack of information. For example, for lymphoma patients who fail in the front-line chemotherapy, there are many salvage regimens available that can form multiple AIs. However, it is difficult to select a pair of AIs for sample size calculation because these regimens are equally likely to be effective. In such situations, powering a SMART based on an omnibus test is more reasonable than based on a pairwise comparison. Ogabagaber et al. (2016) recently gave an excellent review of sample size calculation for SMARTs, and considered sample size determination for a Wald test based on inverse probability weighted estimation under three specific SMART designs. In this Chapter, I will proposed a gate-keeping approach using an omnibus Wald test based on maximum likelihood estimation, and discuss the method of calculating sample size associated with this test in Chapter 2.2.

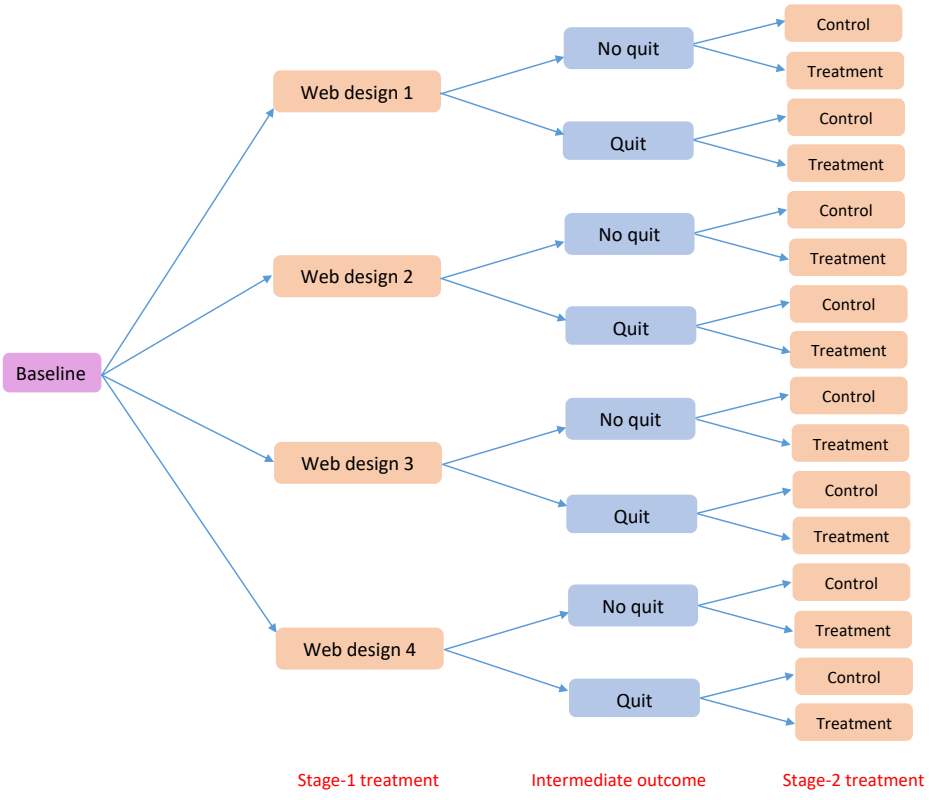


Figure 2.1. Design structure of smoking cessation trial

While there is a wide applications of SMART, we will illustrate the proposed method using data extracted from a smoking cessation randomized trial at the Center for Health Communications Research (CHCR) at the University of Michigan (Strecher et al., 2008). Figure 2.1 is the design structure of this trial. Briefly, in this trial, four options of behavioral treatments defined by two 2-level factors were available for randomization at the first treatment stage, and then two options at the second stage. The study aimed to improve adult smoking quitting rate. The AIs embedded in this example are listed in Tables 2.1 along with some numerical results of the analysis. Additional details are given in Chapter 2.4.

Table 2.1. Multiple Comparison of adaptive interventions embedded in CHCR study. The P values were obtained from pairwise tests comparing each AI with the observed best ($g = 1$). Bonferroni's adjustment would require $P < 0.0004$ to achieve an overall significance at 5%.

	Stage-1	Stage-2 Treatment for			
AI (g)	Treatment	Non-response	Response	$\hat{\theta}_g$ (sd)	P-value
1	Low source/low depth	control	control	0.43 (0.10)	-
2	Low source/low depth	control	treatment	0.42 (0.08)	0.913
3	Low source/low depth	treatment	control	0.31 (0.08)	0.057
4	Low source/low depth	treatment	treatment	0.30 (0.06)	0.194
5	Low source/high depth	control	control	0.42 (0.08)	0.962
6	Low source/high depth	control	treatment	0.27 (0.07)	0.191
7	Low source/high depth	treatment	control	0.35 (0.07)	0.525
8	Low source/high depth	treatment	treatment	0.20 (0.06)	0.044
9	High source/low depth	control	control	0.22 (0.08)	0.096
10	High source/low depth	control	treatment	0.32 (0.06)	0.363
11	High source/low depth	treatment	control	0.29 (0.08)	0.290
12	High source/low depth	treatment	treatment	0.40 (0.06)	0.787
13	High source/high depth	control	control	0.40 (0.08)	0.841
14	High source/high depth	control	treatment	0.37 (0.07)	0.630
15	High source/high depth	treatment	control	0.42 (0.07)	0.965
16	High source/high depth	treatment	treatment	0.39 (0.06)	0.741

sd: estimated asymptotic standard deviation of $\hat{\theta}_g$

The rest of this chapter is organized as follows. Chapter 2.2 specifies setting, notation, and model, introduces the proposed Wald test, and discusses sample size determination based on the proposed test. Chapter 2.3 evaluates finite sample performance of the proposed method using simulation. Chapter 2.4 shows an application using the real data example of the smoking cessation trial. It ends with some discussion in Chapter 2.5.

2.2 An Omnibus Test for Comparing Multiple Adaptive Interventions

2.2.1 Setting, Notation and Model

For brevity in exposition, I consider SMART designs with the primary objective to compare two-stage AIs, although the notation can be readily extended to any SMART design with more than 2 stages. Suppose that there are I treatment options $T_1, \ldots, T_I$ at Stage 1, and under treatment T_i, there are J_i possible intermediate responses, denoted by $R_{i1}, \ldots, R_{iJ_i}$ for $i = 1, \ldots, I$. Next suppose that for a subject who receives treatment T_i at Stage 1 and has an intermediate response of R_{ij}, there are K_{ij} treatment options at Stage 2, namely $S_{ij1}, \ldots, S_{ijK_{ij}}$, for $i = 1, \ldots, I$; $j = 1, \ldots, J_i$. The design structure of a two-stage SMART is thus completely specified by the set of

$$\{(T_i, R_{ij}, S_{ijk}) : \ i = 1, \ldots, I; \ j = 1, \ldots, J_i; \ k = 1, \ldots, K_{ij}\}$$

and is depicted in Figure 2.1. Let U_l denote the treatment received at Stage 1, X_l the intermediate response observed at the end of Stage 1, V_l the treatment received at Stage 2, and Y_l the final primary outcome observed at the end of Stage 2 for subject l, where $l = 1, \ldots, n$. For example, the primary outcome Y_l is a binary indicator of

quitting smoking in the CHCR smoking cessation study. Let $\pi_i = \Pr(U_l = T_i)$ be the randomization probability of assigning T_i to subject l at Stage 1, and $\pi_{ijk} = \Pr(V_l = S_{ijk}|U_l = T_i, X_l = R_{ij})$ be the randomization probability of assigning treatment S_{ijk} to patient l given the clinical history of stage-1 treatment and response $(U_l = T_i, X_l = R_{ij})$. The randomization scheme of a two-stage SMART can be completely specified by

$$\{(\pi_i, \pi_{ijk}) : \ i = 1, \ldots, I; \ j = 1, \ldots, J_i; \ k = 1, \ldots, K_{ij}\},$$

and is depicted in Figure 2.2. Here I subtracted the subject indicator, l, for simplicity.

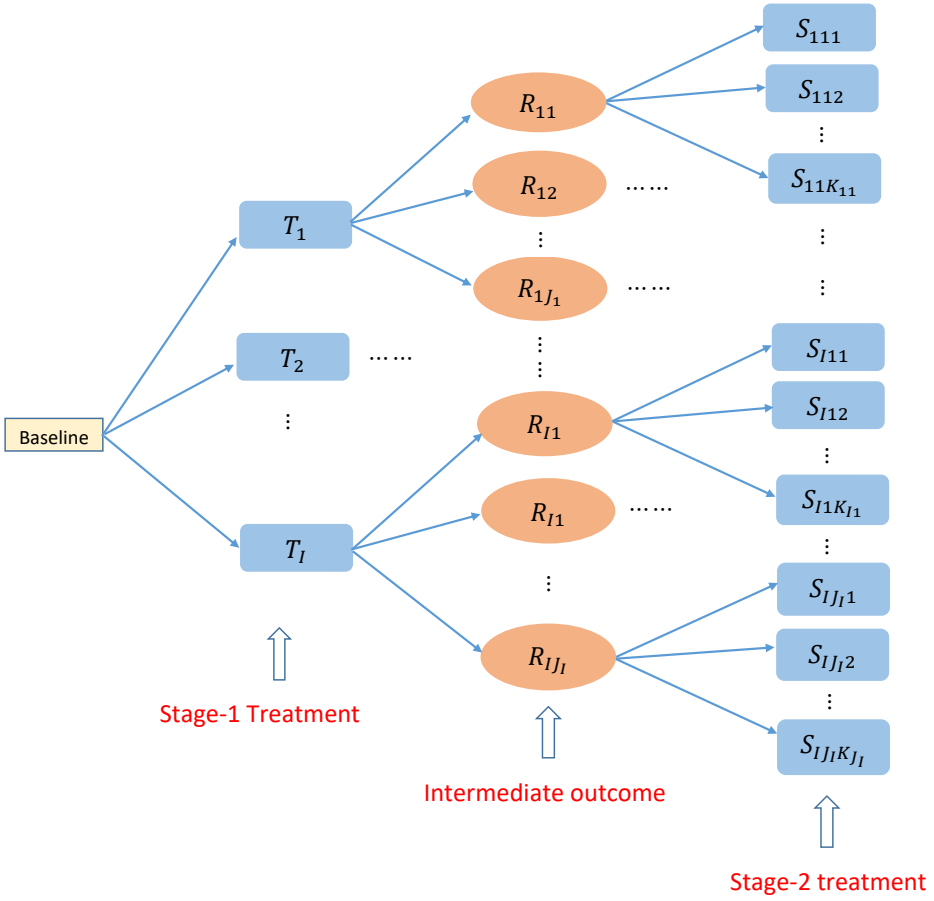


Figure 2.2. General scheme of two-stage SMART design

The data obtained from the lth patient who has completed a SMART can be summarized as (U_l, X_l, V_l, Y_l) for $l = 1, \ldots, n$, and are assumed to be independent and identically with the following distributions:

$$\Pr(U_l = T_i) = \pi_i, \quad i = 1, \ldots, I,$$

$$\Pr(X_l = R_{ij} | U_l = T_i) = p_{ij}, \quad j = 1, \ldots, J_i, \ i = 1, \ldots, I,$$

$$\Pr(V_l = S_{ijk} | U_l = T_i, X_l = R_{ij}) = \pi_{ijk}, \quad k = 1, \ldots, K_{ij}, \ j = 1, \ldots, J_i, \ i = 1, \ldots, I,$$

$$Y_l | (U_l = T_i, X_l = R_{ij}, V_l = S_{ijk}) \sim f\big(y_l | \phi_{ijk}, \tau_{ijk}\big),$$

where p_{ij} is the probability of observing intermediate response R_{ij} given stage-1 treatment T_i, ϕ_{ijk} is the parameter of interest, and τ_{ijk}, possibly a vector, is the nuisance parameter. I assume that $f\big(y_l | \phi_{ijk}, \tau_{ijk}\big)$ satisfies the regularity conditions specified in Theorem 5.39 in van der Varrt (1998) which guarantee the asymptotic efficiency of the maximum likelihood estimator (MLE) of (ϕ_{ijk}, τ_{ijk}). For example, data arising from a distribution belonging to the exponential family will satisfy the conditions. The stage-specific randomization probabilities π_i and π_{ijk} are known by design in SMART.

An two-stage AI, under which a patient receives treatment T_i at Stage 1 and will receive treatment $S_{ijk_{ij}}$ at Stage 2 if an intermediate response of R_{ij} is observed, can be denoted by

$$d_{i;k_{i1},\ldots,k_{iJ_i}} = (T_i; S_{i1k_{i1}}, \ldots, S_{iJ_ik_{iJ_i}}),$$

where $i = 1, \ldots, I$ and $k_{ij} = 1, \ldots, K_{ij_i}$ for $j_i = 1, \ldots, J_i$. Note that $(k_{i1}, \ldots, k_{iJ_i})$ is an element in the product set $\prod_{j=1}^{J_i} \{1, \ldots, K_{ij}\}$. Under the general SMART design given in Figure 2.2, the total number of AIs with T_i at Stage 1 is $G_i = \prod_{j=1}^{J_i} K_{ij}$, and thus the

total number of AIs embedded in a SMART is $G = \sum_{i=1}^{I} G_i$.

Let $\theta_{i;k_{i1},\ldots,k_{iJ_i}}$ be the value of an two-stage AI embedded in a general SMART design described in previous paragraph. Note that under an AI, $d_{i;k_{i1},\ldots,k_{iJ_i}}$, there are J_i possible treatment sequences, (T_i, R_{ij}, S_{ijk}), which a patient can possibly follow. Therefore, the AI value can be defined as the expected outcome Y across all the possible treatment sequences under this AI, and thus be evaluated as

$$\theta_{i;k_{i1},\ldots,k_{iJ_i}} = \sum_{j=1}^{J_i} p_{ij}\phi_{ijk_{ij}},$$

An AI $d_{i;k_{i1},\ldots,k_{iJ_i}}$ is said to be the best among all the AIs embedded in a SMART if it has the same value as $d_{i^*;k^*_{i^*1},\ldots,k^*_{i^*J_{i^*}}}$ where

$$\left\{d_{i^*;k^*_{i^*1},\ldots,k^*_{i^*J_{i^*}}}\right\} \in \text{argmax}_{1\le i\le I; j=1,\ldots,J_i; 1\le k_{ij}\le K_{ij}} \theta_{i;k_{i1},\ldots,k_{iJ_i}}.$$

Note that there could be more than one best AI embedded in a SMART design.

2.2.2 Maximum Likelihood Estimation

In this section, I consider the maximum likelihood estimator (MLE) for an AI value $\theta_{i;k_{i1},\ldots,k_{iJ_i}}$ and examine its asymptotic distribution. The MLE of an AI is obtained by plugging in the MLEs of p_{ij}'s and ϕ_{ijk}'s, which are obtained by maximizing the joint distribution of $\big\{(U_l, X_l, V_l, Y_l);\ l = 1, \ldots, n\big\}$.

Specifically, to obtain a MLE, I first derive the log-likelihood function based on the

joint probability distribution of $\{(U_l, X_l, V_l, Y_l);\ l = 1, \ldots, n\}$ as

$$\begin{aligned}\log L(p_{ij}, \phi_{ijk}, \tau_{ijk}) = &\sum_{l=1}^{n}\sum_{i=1}^{I}\sum_{j=1}^{J_i}\sum_{k=1}^{K_{ij}} I(U_l = T_i, X_l = R_{ij}, V_l = S_{ijk}) \log f(y_l|\phi_{ijk}, \tau_{ijk}) \\ &+ \sum_{l=1}^{n}\sum_{i=1}^{I}\sum_{j=1}^{J_i}\sum_{k=1}^{K_{ij}} I(U_l = T_i, X_l = R_{ij}, V_l = S_{ijk}) \log p_{ij} + \text{constant}.\end{aligned}$$

And then derive the score functions as

$$\begin{aligned}\frac{\partial \log L}{\partial p_{ij}} &= \frac{\sum_{l=1}^{n}\sum_{k=1}^{K_{ij}} I(U_l = T_i, X_l = R_{ij}, V_l = S_{ijk})}{p_{ij}} \\ &\quad - \frac{\sum_{l=1}^{n}\sum_{j'\neq j}\sum_{k=1}^{K_{ij'}} I(U_l = T_i, X_l = R_{ij'}, V_l = S_{ij'k})}{1 - p_{ij}}, \\ \frac{\partial \log L}{\partial \phi_{ijk}} &= \sum_{l=1}^{n} I(U_l = T_i, X_l = R_{ij}, V_l = S_{ijk}) \frac{\partial \log f(y_l|\phi_{ijk}, \tau_{ijk})}{\partial \phi_{ijk}}, \\ \frac{\partial \log L}{\partial \tau_{ijk}} &= \sum_{l=1}^{n} I(U_l = T_i, X_l = R_{ij}, V_l = S_{ijk}) \frac{\partial \log f(y_l|\phi_{ijk}, \tau_{ijk})}{\partial \tau_{ijk}},\end{aligned}$$

where $i = 1, \ldots, I$, $j = 1, \ldots, J_i$, and $k = 1, \ldots, K_{ij}$.

Thus, the MLE for p_{ij} is

$$\hat{p}_{ij} = \frac{\sum_{l=1}^{n}\sum_{k=1}^{K_{ij}} I(U_l = T_i, X_l = R_{ij}, V_l = S_{ijk})}{\sum_{l=1}^{n}\sum_{j'=1}^{J_i}\sum_{k=1}^{K_{ij'}} I(U_l = T_i, X_l = R_{ij'}, V_l = S_{ij'k})},$$

and the MLEs for ϕ_{ijk} and τ_{ijk} are the ones based on the subset of Y_l's such that $U_l = T_i, X_l = R_{ij}, V_l = S_{ijk}$, and are denoted as $\hat{\phi}_{ijk}$ and $\hat{\tau}_{ijk}$, respectively. For example, in a SMART that we are interested in the continuous primary outcome that is assumed to follow a normal distribution,

$$Y_l|(U_l = T_i, X_l = R_{ij}, V_l = S_{ijk}) \sim N(\phi_{ijk}, \sigma^2_{ijk}),$$

where the mean, ϕ_{ijk}, is the parameter of interest and the variance, σ^2_{ijk}, is the nuisance parameter. I can get the MLE for ϕ_{ijk} and σ^2_{ijk} by taking the sample average and variance of those patients with the treatment history of (T_i, R_{ij}, S_{ijk}). Note that in general there is no closed form expression for $\hat{\phi}_{ijk}$ or $\hat{\tau}_{ijk}$.

For each i, denote

$$\boldsymbol{p}_i = \begin{pmatrix} p_{i1} \\ \vdots \\ p_{iJ_i} \end{pmatrix}, \quad \boldsymbol{\phi}_{ij} = \begin{pmatrix} \phi_{ij1} \\ \vdots \\ \phi_{ijK_{ij}} \end{pmatrix}, \quad \boldsymbol{\phi}_i = \begin{pmatrix} \boldsymbol{\phi}_{i1} \\ \vdots \\ \boldsymbol{\phi}_{iJ_i} \end{pmatrix},$$

so that $\hat{\boldsymbol{p}}_i, \hat{\boldsymbol{\phi}}_{ij}$ and $\hat{\boldsymbol{\phi}}_i$ their MLEs, respectively.

Let $\boldsymbol{\theta}_i$ be the vector of values of all the AIs starting with T_i at Stage 1, which are denoted by $\theta_{i,k_{i1},\ldots,k_{iJ_i}}$ and arranged in the lexicographical order of $(k_{i1}, \ldots, k_{iJ_i})$. Also, let $G_i = \prod_{j=1}^{J_i} K_{ij}$ and $m_i = \sum_{j=1}^{J_i} K_{ij}$. Here $\boldsymbol{\theta}_i$ can be expressed in two forms as

$$\boldsymbol{\theta}_i = \boldsymbol{A}_i \boldsymbol{\Lambda}_i(\boldsymbol{p}_i)\boldsymbol{\phi}_i = \boldsymbol{A}_i \boldsymbol{\Gamma}_i(\boldsymbol{\phi}_i)\boldsymbol{p}_i \ , \tag{1}$$

where $\boldsymbol{A}_i$ is an $G_i \times m_i$ matrix defined as

$$\boldsymbol{A}_i = \big(\boldsymbol{I}_{K_{i1}} \otimes \mathbf{1}_{K_{i2}} \otimes \cdots \otimes \mathbf{1}_{K_{iJ_i}} | \mathbf{1}_{K_{i1}} \otimes \boldsymbol{I}_{K_{i2}} \otimes \cdots \otimes \mathbf{1}_{K_{iJ_i}} | \cdots | \mathbf{1}_{K_{i1}} \otimes \cdots \otimes \mathbf{1}_{K_{i(J_i-1)}} \otimes \boldsymbol{I}_{K_{iJ_i}}\big), \tag{2}$$

$\otimes$ denotes the Kronecker product, and $\boldsymbol{\Lambda}_i(\boldsymbol{p}_i)$ is an $m_i \times m_i$ block diagonal matrix

$$\boldsymbol{\Lambda}_i(\boldsymbol{p}_i) = \mathrm{bdiag}\big\{p_{ij}\boldsymbol{I}_{K_{ij}};\ j = 1, \ldots, J_i\big\},$$

and $\boldsymbol{\Gamma}_i(\boldsymbol{\phi})$ is an $G_i \times J_i$ block diagonal matrix

$$\boldsymbol{\Gamma}_i(\boldsymbol{\phi}_i) = \text{bdiag}\big\{\boldsymbol{\phi}_{ij};\ j = 1, \ldots, J_i\big\},$$

where "bdiag$\{\cdot\}$" denotes block diagonal matrix.

The two expressions of the MLE of $\boldsymbol{\theta}_i$ in (1) can be respectively expressed as

$$\hat{\boldsymbol{\theta}}_i = \boldsymbol{A}_i \boldsymbol{\Lambda}(\hat{\boldsymbol{p}}_i)\hat{\boldsymbol{\phi}}_i = \boldsymbol{A}_i \boldsymbol{\Gamma}_i(\hat{\boldsymbol{\phi}}_i)\hat{\boldsymbol{p}}_i.$$

Now define

$$\begin{aligned}
&\boldsymbol{\Sigma}_{\boldsymbol{p}_i} = \pi_i^{-1}\left(\text{diag}\{\boldsymbol{p}_i\} - \boldsymbol{p}_i\boldsymbol{p}_i^T\right),\\
&\boldsymbol{\Sigma}_{\boldsymbol{\phi}_i} = \text{bdiag}\big\{\boldsymbol{\Sigma}_{\phi_{i1}}, \ldots, \boldsymbol{\Sigma}_{\phi_{iJ_i}}\big\},\\
&\boldsymbol{\Sigma}_{\boldsymbol{\phi}_{ij}} = big(\pi_i p_{ij}\pi_{ijk}\big)^{-1}\text{bdiag}\big\{\sigma^2(\boldsymbol{\phi}_{ijk}, \boldsymbol{\tau}_{ijk});\ k = 1, \ldots, K_{ij}\big\},\\
\text{and}\ \ &\sigma^2(\phi_{ijk}, \tau_{ijk}) = \left(\boldsymbol{\imath}_{\phi_{ijk}\phi_{ijk}} - \boldsymbol{\imath}^T_{\phi_{ijk}\tau_{ijk}} \cdot \boldsymbol{\imath}^{-1}_{\tau_{ijk}\tau_{ijk}} \cdot \boldsymbol{\imath}_{\phi_{ijk}\tau_{ijk}}\right)^{-1},
\end{aligned}$$

where

$$\begin{pmatrix} \boldsymbol{\imath}_{\phi_{ijk}\phi_{ijk}} & \boldsymbol{\imath}^T_{\phi_{ijk}\tau_{ijk}} \\ \boldsymbol{\imath}_{\phi_{ijk}\tau_{ijk}} & \boldsymbol{\imath}_{\tau_{ijk}\tau_{ijk}} \end{pmatrix}$$

is the block Fisher's information matrix of distribution $f(y|\phi_{ijk}, \tau_{ijk})$.

Theorem 1 *Let $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1^T, \ldots, \boldsymbol{\theta}_I^T)^T$. Under regularity conditions given in Theorem 5.39 in van der Vaart (1998), as $n \to \infty$,*

$$\sqrt{n}(\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}) \xrightarrow{d} N\big(\mathbf{0},\ \boldsymbol{\Sigma}\big), \tag{3}$$

where $\boldsymbol{\Sigma} = \text{bdiag}\{\boldsymbol{\Sigma}_{\boldsymbol{\theta}_1}, \ldots, \boldsymbol{\Sigma}_{\boldsymbol{\theta}_I}\}$, $\boldsymbol{\Sigma}_{\theta_i} = \boldsymbol{A}_i\big(\boldsymbol{\Gamma}_i(\boldsymbol{\phi}_i)\boldsymbol{\Sigma}_{\boldsymbol{p}_i}\boldsymbol{\Gamma}_i(\boldsymbol{\phi}_i)^T + \boldsymbol{\Lambda}_i(\boldsymbol{p}_i)\boldsymbol{\Sigma}_{\boldsymbol{\phi}_i}\boldsymbol{\Lambda}_i(\boldsymbol{p}_i)\big)\boldsymbol{A}_i^T$, *and*

$$\text{rank}\left(\boldsymbol{\Sigma}_{\boldsymbol{\theta}_i}\right) = \sum_{j=1}^{J_i} K_{ij} - J_i + 1 = m_i - J_i + 1. \tag{4}$$

Theorem 1 shows that the MLE $\hat{\boldsymbol{\theta}}_i$ is asymptotically normal, and that the asymptotic covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\theta}_i}$ is not of full rank G_i when $J_i \geq 2$. This is the fundamental distributional result on which all the proposed techniques are built. The proof of Theorem 1 is given in the Appendix 1.

2.2.3 Wald Test and Sample Size Determination

I propose in this section a Wald-type omnibus test and discuss the sample size determination based on the proposed test. For ease of exposition, I use θ_g to denote the gth component of $\boldsymbol{\Theta}$, where $g = 1, \ldots, G$, G is the total number of AIs and $\boldsymbol{\Theta}$ is the vector of all the AI values, $\theta_{i;k_{i1},\ldots,k_{iJ_i}}$, embedded in a SMART, arranged in a lexicographical order of $(i; k_{i1}, \cdots, k_{iJ_i})$. As an initial step of a gate-keep approach to identify the best AI, an omnibus test of equality is considered with the following hypotheses:

$$H_0 : \theta_1 = \cdots = \theta_G \text{ versus } H_1\text{: } \theta_g\text{'s are not all equal for } g = 1, \ldots, G. \tag{5}$$

Let $\boldsymbol{C} = (\boldsymbol{1}_{G-1}| - \boldsymbol{I}_{G-1})$ be a $(G-1) \times G$ contrast matrix such that the first column is a $(G-1)$ vector of 1's and the j-th column is a $(G-1)$ vector whose $(j-1)$-th entry is -1 and other entries are zeros, where $2 \leq j \leq G$. Let $\boldsymbol{\Sigma}$ be the covariance matrix of $\hat{\boldsymbol{\Theta}}$

defined in Theorem 1, and $\hat{\boldsymbol{\Sigma}}$ is the plug-in estimator of $\boldsymbol{\Sigma}$ by replacing $p_{ij}, \theta_{ijk}, \tau_{ijk}$ with their MLEs $\hat{p}_{ij}, \hat{\theta}_{ijk}, \hat{\tau}_{ijk}$, respectively. Then a Wald-type test statistic can be written as

$$Q = n(\boldsymbol{C}\hat{\boldsymbol{\Theta}})^T(\boldsymbol{C}\hat{\boldsymbol{\Sigma}}\boldsymbol{C}^T)^-(\boldsymbol{C}\hat{\boldsymbol{\Theta}}), \tag{6}$$

where $\boldsymbol{M}^-$ denotes the generalized inverse of a square matrix $\boldsymbol{M}$.

Theorem 2 *Suppose all regularity conditions given in Theorem 5.39 in van der Vaart (1998) hold. Under H_0 in (5) and as $n \to \infty$, $Q \xrightarrow{d} \chi^2_\nu$, where*

$$\nu = \sum_{i=1}^{I}\sum_{j=1}^{J_i} K_{ij} - \sum_{i=1}^{I} J_i + I - 1 = \sum_{i=1}^{I} m_i - \sum_{i=1}^{I} J_i + I - 1. \tag{7}$$

In addition, under a sequence of local alternatives $\{\boldsymbol{\Theta}_n\}$ such that

$$\lim_{n\to\infty} n(\boldsymbol{C}\boldsymbol{\Theta}_n)^T(\boldsymbol{C}\boldsymbol{\Sigma}\boldsymbol{C}^T)^-(\boldsymbol{C}\boldsymbol{\Theta}_n) = \lambda^* > 0, \tag{8}$$

$Q \xrightarrow{d} \chi^2_\nu(\lambda^*)$, a noncentral chi-squared distribution of ν degrees of freedom with noncentrality parameter λ^*.

As a consequence to Theorem 2, an asymptotic level α test rejects H_0 in (5) if the test statistic $Q > \chi^2_{\nu,\alpha}$, where $\chi^2_{\nu,\alpha}$the $(1-\alpha)$th percentile of the central chi-squared distribution with ν degrees of freedom. Note that in a special case that $\mathrm{rank}(\Sigma_{\theta_i}) = 1$, it reduces to the regular Wald test for comparing non-adaptive intervention sequences.

Theorem 2 also provides a basis for sample size determination, which may proceed prescriptively as follows:

Step 1: For a given design structure of SMART $\{T_i, R_{ij}, S_{ijk}\}$, where $i = 1, \ldots, I; j = 1, \ldots, J_i; k = 1, \ldots, K_{ij}$, calculate the degrees of freedom ν according to (7).

Step 2: For a prespecified type I error rate α and a targeted power of $1-\beta$, determine the noncentrality parameter λ^* required under the alternative hypothesis by solving $\chi^2_{\nu,1-\beta}(\lambda^*) = \chi^2_{\nu,\alpha}(0)$, where $\chi^2_{\nu,\alpha}(\lambda^*)$ denotes the $(1-\alpha)$th percentile of a noncentral chi-squared distribution with degrees of freedom ν and λ^* is the noncentrality parameter.

Step 3: For given design parameters $\{\pi_i, \pi_{ijk}\}$, assumed intermediate response probabilities $\{p_{ij}\}$ and primary outcome parameter values $\{\phi_{ijk}, \tau_{ijk}\}$ for the outcome distribution f, calculate the targeted AI values $\boldsymbol{\Theta}^*$ and its covariance $\boldsymbol{\Sigma}^*$, so that standardized overall effect size can be calculated according to (9):

$$\Delta = \left(\boldsymbol{C}\boldsymbol{\Theta}^*\right)^T \left(\boldsymbol{C}\boldsymbol{\Sigma}^*\boldsymbol{C}^T\right)^- \left(\boldsymbol{C}\boldsymbol{\Theta}^*\right). \tag{9}$$

Step 4: The total number of patients required for the study is

$$n = \frac{\lambda^*}{\Delta}.$$

Table 2.2 gives the values of noncentrality parameter λ^* under some commonly used targeted type I and II errors, (α, β), in clinical trials. Generally, smaller error rates and larger degrees of freedom, which reflects the numbers of treatment options and response categories, require a larger λ^*, and hence a larger sample size per Step 4 above.

Table 2.2. Noncentrality parameter (λ^*) of chi-squared distribution with ν degrees of freedom. α is type I error, β is type II error.

	$(\alpha,\ \beta)$					
Degrees of freedom (ν)	(0.01, 0.10)	(0.01, 0.20)	(0.05, 0.10)	(0.05, 0.20)	(0.10, 0.10)	(0.10, 0.20)
2	17.42	13.88	12.65	9.63	10.45	7.71
3	19.24	15.45	14.17	10.90	11.79	8.80
4	20.73	16.75	15.41	11.94	12.88	9.68
5	22.02	17.87	16.47	12.83	13.81	10.44
6	23.18	18.87	17.42	13.62	14.65	11.13
7	24.23	19.78	18.28	14.35	15.41	11.75
8	25.20	20.63	19.08	15.02	16.11	12.32
9	26.12	21.42	19.81	15.65	16.76	12.86
10	26.98	22.17	20.53	16.24	17.38	13.36
11	27.79	22.88	21.20	16.80	17.96	13.84
12	28.57	23.56	21.83	17.34	18.52	14.30
13	29.31	24.21	22.44	17.85	19.05	14.74
14	30.03	24.83	23.02	18.34	19.56	15.16
15	30.71	25.43	23.58	18.81	20.06	15.56
16	31.38	26.01	24.13	19.27	20.53	15.95
17	32.02	26.57	24.65	19.71	20.99	16.33
18	32.65	27.11	25.16	20.14	21.43	16.69
19	33.25	27.64	25.65	20.56	21.87	17.05
20	33.84	28.16	26.13	20.96	22.29	17.39

2.3 Finite Sample Performances

Having established the asymptotic properties of the Wald test in the previous section, I evaluate its performances in finite sample size settings using simulation in this section. In addition, the test is applied as a gate-keeping method: if the test fails to reject H_0 in (5), I will stop further comparison and conclude that there is no sufficient evidence to support any AI being better than the others. However, if H_0 is rejected, I proceed to select the AI with the highest estimated value and recommend it for further clinical evaluation.

The properties of the test and the selection procedure are examined under a variety of SMART designs (Chapter 2.3.1) and outcome scenarios (Chapter 2.3.2).

2.3.1 SMART Designs

Figure 2.3 gives the structures of three two-stage SMART designs considered in simulation. The first design structure (DS1) mimics the situation in which there are two treatment options at each decision making point, that is, $T_i, S_{ijk} \in \{0, 1\}$, and binary intermediate outcome, that is, $R_{ij} \in \{0, 1\}$ for $i, j, k = 1, 2$. As a result, there are eight possible AIs embedded in DS1. Under DS2 and DS3, there are also two treatment options at Stage 1. However, randomization at Stage 2 may be restricted for patients with certain intermediate responses; and hence there are 4 and 3 embedded AIs in DS2 and DS3, respectively.

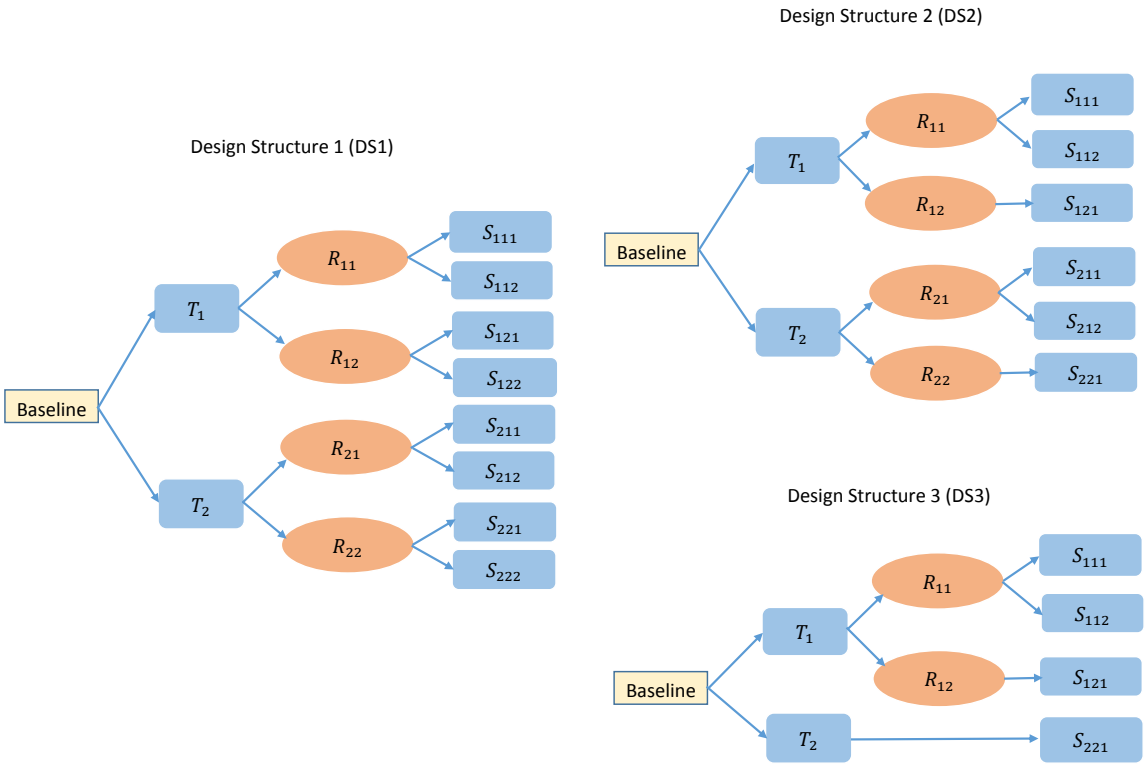


Figure 2.3. Design structures considered in the simulation.

Under each design structure, a SMART design is completely specified by the set $\{\pi_i, \pi_{ijk}\}$ of randomization probabilities defined in Chapter 2.2.1. I considered three sets of randomization probabilities for each design structure in Figure 2.2. First, I considered balanced randomization (BR) scheme, that is, $\Pr(U = 1) = 0.5$ and $\Pr(V = 1|U, X) = 0.5$ whenever there is an option of randomization at Stage 2. Second, I considered an unbalanced randomization (UBR) scheme, where $\Pr(U = 1) = 0.7$ and $\Pr(V = 1|U, X) = 0.7$ whenever there is an option of stage-2 randomization. Third, I considered $\Pr(U = 1) = 0.5$ at Stage 1, $\Pr(V = U|U, X = 0) = 0.3$ and $\Pr(V = U|U, X = 1) = 0.7$ at Stage 2, whenever there is an option of second stage randomization. Under this scheme, Stage 2 implements a randomized play-the-winner (RPTW) rule for the situations where the first and the second stage treatment options are identical.

In summary, the three design structures (DS1, DS2, DS3) and the three randomization schemes (BR, UBR, RPTW) yielded 9 SMART designs under which the multiple comparison procedures were evaluated in the simulation.

2.3.2 Outcome Scenarios in Simulation

In a simulated SMART trial with a total sample size of n, the treatment assignment (U_l, V_l) of the lth patient was generated according to one of the randomization schemes in Chapter 2.3.1. The intermediate response rate was set as $\Pr(X_l = 1|U_l = T_i) = 1/3$ for $T_i \in \{0, 1\}$. Given the lth subject's treatment history and intermediate response (T_i, R_{ij}, S_{ijk}), his or her outcome Y_l was randomly generated from a normal distribution with mean $\phi_{ijk} = \phi(T_i, R_{ij}, S_{ijk})$ and variance $\sigma^2 = 100$, where the conditional mean ϕ_{ijk}

was specified by

$$\phi(T_i, R_{ij}, S_{ijk}) = \beta_0 + \beta_1 T_i + \beta_2 R_{ij} + \beta_3 S_{ijk} + \beta_4 T_i R_{ij} + \beta_5 T_i S_{ijk} + \beta_6 R_{ij} S_{ijk} + \beta_7 T_i R_{ij} S_{ijk} \quad (10)$$

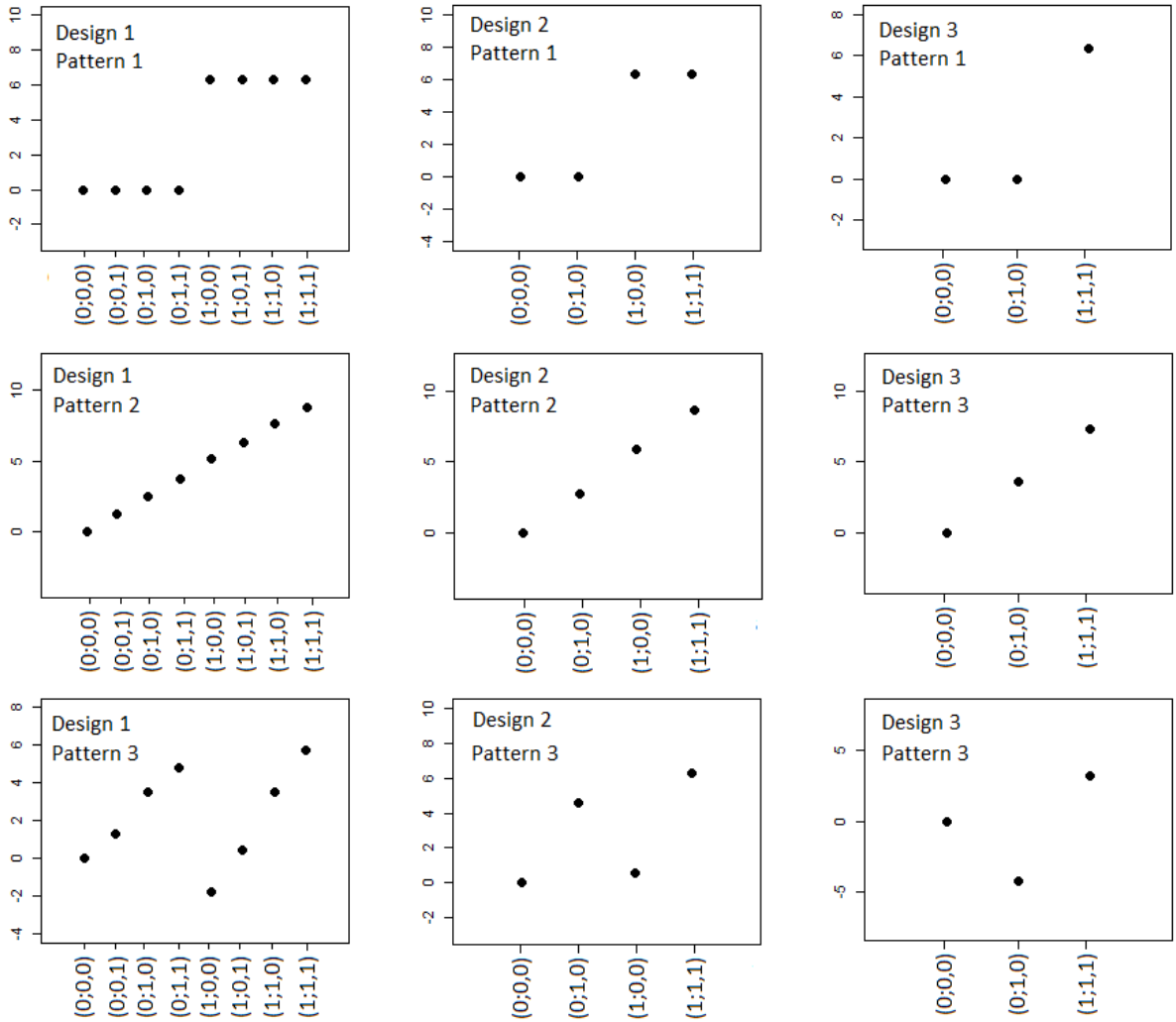
for $T_i, R_{ij}, S_{ijk} \in \{0, 1\}$. The parameter $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \beta_7)^T$ was chosen so that the true values $\theta_{i;k_{i1},\ldots,k_{iJ_i}}$'s would follow the patterns displayed in Figure 2.4. Under Value Pattern 1 (VP1; top panel), AIs with the same stage-1 treatment had the same values; under VP2, the values of the AIs were uniformly higher if their stage-1 treatment was $U = 1$; under VP3 (bottom panel), the best AI had stage-1 treatment $U = 1$ while the second best had stage-1 treatment $U = 0$, and so on and so forth, following an alternating pattern. The value of $\boldsymbol{\beta}$ was chosen so that the effect size was $\Delta = 0.05$ or 0.10. For example, under VP1, $\beta_0 = \beta_2 = \cdots = \beta_7 = 0$ and $\beta_1 = 4.48$ and 6.33 yielded $\Delta = 0.05$ and 0.10, respectively. Details about how to choose $\boldsymbol{\beta}$ for each pattern are provided in Appendix 2. All sets of $\boldsymbol{\beta}$ values used in the simulation scenarios are given in Table 2.3.

Table 2.3. Values of $(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \beta_7)$ used in simulations.

		$(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \beta_7)$	
Design Structure	Value Pattern	$\Delta = 0.05$	$\Delta = 0.10$
DS1	VP1	(0, 4.48, 0, 0, 0, 0, 0, 0)	(0, 6.33, 0, 0, 0, 0, 0, 0)
DS1	VP2	(0, 3.63, 0, 2.62, 0, 0, 0, 0)	(0, 5.13, 0, 3.70, 0, 0, 0, 0)
DS1	VP3	(0, 1.86, 0, 3.73, -9.32, 1.86, -0.93, 0)	(0, 2.64, 0, 5.82, -13.20, 2.64, -1.32, 0)
DS2	VP1	(0, 4.48, 0, 0, 0, 0, 0, 0)	(0, 6.33, 0, 0, 0, 0, 0, 0)
DS2	VP2	(0, 0, 0, 2.88, 12, 0, 0, 0)	(0, 0, 0, 4.13, 17.70, 0, 0, 0)
DS2	VP3	(0, -1.21, 0, 4.82, 4.82, 1.21, 0, 0)	(0, -1.72, 0, 6.87, 6.87, 1.72, 0, 0)
DS3	VP1	(0, 4.48, 0, 0, 0, 0, 0, 0)	(0, 6.33, 0, 0, 0, 0, 0, 0)
DS3	VP2	(0, 1.29, 0, 3.88, 0, 0, 0, 0)	(0, 1.82, 0, 5.47, 0, 0, 0, 0)
DS3	VP3	(0, 0, 0, -4.46, 0, 6.69, 0, 0)	(0, 0, 0, -6.36, 0, 9.54, 0, 0)

Details of design structures (DS) and value patterns (VP) are given in Figures 2.2 and 2.3

Figure 2.4. Value patterns of AIs considered in the simulation.



2.3.3 Gate-Keeping Approach for AI Selection

Here I evaluate a multiple comparison procedure applies the likelihood-based Wald test proposed in Chapter 2.2.3 as a gate-keeping method: if the test fails to reject H_0 in (5), I will stop further comparison and conclude that there is no sufficient evidence to support any AI being better than the others. If H_0 is rejected, I will proceed to select the AI with the highest estimated value and recommend it for further clinical investigation.

Table 2.4 gives the actual type I error rates of the proposed Wald test at 5% nominal significance level, based on 5,000 simulation replicates under the 9 SMART designs described in Chapter 2.3.1 with a total sample size of $n = 200$ in outcomes scenarios generated with $\boldsymbol{\beta} = (0, 0, 0, 0, 0, 0, 0, 0)$ in (10). Overall, the actual type I error rates are very close to the nominal level in all the 9 designs. For comparison purposes, I also considered pairwise testing procedures comparing the AIs, with or without multiplicity adjustment. The procedure without multiplicity adjustment would reject the null H_0 in (5) if *any* pairwise test had a P-value less than 0.05. As expected, this procedure led to inflated type I error rates, especially under DS1 where there were many comparisons (i.e., 28 in total). For the pairwise testing procedure with multiplicity adjustment, we used the Bonferroni's method and adjusted significance level for each individual test according to the number of comparisons under each DS (i.e., 28, 6, and 3 for DS1, DS2, and DS3, respectively). Specifically, we would reject the H_0 in (5) if any pairwise test had a P-value less than 0.0018, 0.0083, and 0.0167 under DS1, DS2, and DS3, respectively. Table 2.4 shows that the Bonferroni's correction was conservative, especially under DS1 where many comparisons were accounted for.

Table 2.4. Type I error rate of the proposed Wald test and the two pairwise test procedures at 5% ominal significance under H_0 in (5) and a total sample size of $n = 200$.

Design structure	Randomization scheme	Wald test	Pairwise tests (no adjustment)	Pairwise tests (Bonferroni's)
DS1	BR	0.051	0.355	0.022
	UBR	0.048	0.293	0.015
	RPTW	0.049	0.336	0.020
DS2	BR	0.051	0.196	0.043
	UBR	0.050	0.171	0.035
	RPTW	0.050	0.188	0.039
DS3	BR	0.051	0.114	0.040
	UBR	0.048	0.112	0.039
	RPTW	0.053	0.115	0.042

Table 2.5 compares the power of the Wald test and those of pairwise tests with Bonferroni's corrections for the outcome scenarios given in Table 2.3 under different SMART designs. Overall, the Wald test was more powerful than the Bonferroni's adjusted pairwise tests in all scenarios and designs considered. For each given effect size (Δ), the power of the pairwise testing procedure had a sharp drop under DS1 when compared with the other design structures, likely due to the needs to adjust for many comparisons. In contrast, while the Wald test also had lower powers under DS1 than under DS2 and DS3, the drop was much less substantial. This demonstrated that an omnibus test was advantageous over a pairwise comparison procedure because the former attenuated the impact of a large number of AIs on the power of a SMART study. In addition, I calculated the theoretical power of the Wald test (cf. Theorem 2) and noted that the asymptotic approximation was accurate when $n = 200$.

Table 2.5. Power of the proposed Wald test and the pairwise tests with Bonferroni's adjustment at 5% overall significant under scenarios given in Table 4 and a total sample size of $n = 200$.

			$\Delta = 0.05$			$\Delta = 0.10$		
Design structure	Value pattern	Randomization scheme	Wald (theo)	Wald (emp)	Pairwise tests (Bonferroni's)	Wald (theo)	Wald (emp)	Pairwise tests (Bonferroni's)
DS1	VP1	BR	0.679	0.672	0.582	0.953	0.951	0.906
		UBR	0.590	0.581	0.405	0.907	0.908	0.753
		RPTW	0.679	0.681	0.478	0.953	0.948	0.847
DS1	VP2	BR	0.679	0.672	0.597	0.952	0.943	0.919
		UBR	0.589	0.573	0.415	0.905	0.908	0.805
		RPTW	0.579	0.570	0.425	0.899	0.892	0.803
DS1	VP3	BR	0.679	0.673	0.470	0.953	0.946	0.833
		UBR	0.669	0.662	0.372	0.949	0.945	0.753
		RPTW	0.620	0.612	0.378	0.925	0.918	0.755
DS2	VP1	BR	0.763	0.761	0.729	0.975	0.977	0.968
		UBR	0.680	0.679	0.577	0.946	0.941	0.888
		RPTW	0.763	0.760	0.677	0.975	0.976	0.944
DS2	VP2	BR	0.763	0.759	0.747	0.975	0.974	0.973
		UBR	0.696	0.691	0.561	0.959	0.958	0.906
		RPTW	0.592	0.589	0.559	0.898	0.900	0.882
DS2	VP3	BR	0.763	0.758	0.648	0.975	0.976	0.935
		UBR	0.726	0.728	0.526	0.964	0.966	0.849
		RPTW	0.673	0.683	0.553	0.941	0.947	0.852
DS3	VP1	BR	0.817	0.808	0.756	0.985	0.985	0.979
		UBR	0.741	0.734	0.649	0.965	0.964	0.928
		RPTW	0.817	0.822	0.769	0.985	0.984	0.970
DS3	VP2	BR	0.817	0.810	0.796	0.985	0.983	0.981
		UBR	0.789	0.781	0.727	0.979	0.977	0.960
		RPTW	0.703	0.701	0.668	0.950	0.946	0.936
DS3	VP3	BR	0.817	0.818	0.802	0.985	0.984	0.981
		UBR	0.561	0.576	0.532	0.867	0.868	0.837
		RPTW	0.878	0.877	0.850	0.995	0.996	0.992

theo: theoretical power; emp: empirical power.

Table 2.6 compares the proposed Wald test based on MLE and the Wald test based on inverse probability weighted estimators (IPWE) described in Ogabagaber et al. (2016). I extracted the scenarios and results of the IPWE Wald test from Table I in Ogabagaber et al. (2016), calculated the sample size required by our proposed test (Chapter 2.2.3), and evaluated the power of our test using simulation. The proposed test generally required a smaller sample size while achieving comparable power with IPWE test. This is because my reference distribution, derived based on asymptotic theory of the MLEs, accounts for the fact that the asymptotic covariance matrix $\mathbf{\Sigma}$ in (3) is generally less than full rank.

Table 2.6. Comparison of the proposed Wald test and the IPWE Wald test.

				Required sample size		Empirical power	
$\Pr(X=1\|U=0)$	$\Pr(X=1\|U=1)$	$\Pr(V=1\|X=1)$	Nominal Power	IPWE	Proposed	IPWE	Proposed
0.5	0.5	0.5	0.80	70	63	0.84	0.84
0.5	0.5	0.7	0.80	79	70	0.85	0.83
0.5	0.5	0.5	0.90	89	80	0.92	0.95
0.5	0.5	0.8	0.90	120	107	0.92	0.95
0.5	0.2	0.5	0.80	83	75	0.83	0.82
0.5	0.2	0.7	0.80	92	81	0.83	0.84
0.5	0.2	0.5	0.90	106	96	0.90	0.94
0.5	0.2	0.8	0.90	134	117	0.92	0.94
0.7	0.5	0.5	0.80	62	56	0.85	0.84
0.7	0.5	0.7	0.80	71	63	0.85	0.85
0.7	0.5	0.5	0.90	79	71	0.92	0.95
0.7	0.5	0.7	0.90	91	81	0.92	0.94
0.2	0.7	0.5	0.80	72	65	0.84	0.83
0.2	0.7	0.7	0.80	82	70	0.84	0.84
0.2	0.7	0.5	0.90	92	83	0.91	0.94
0.2	0.7	0.7	0.90	104	90	0.92	0.92

Table 2.7 gives the selection properties of the gate-keeping method with the proposed Wald test under balanced randomization. As expected, the actual type I error rate, being close to the nominal level of 5%, was equally distributed among all the AIs. Also, AIs with higher true values were selected more often, and the selection accuracy improved as the effect size Δ became larger. Under VP1 where an AI had either a value of 0 or a positive value, the probability of selecting an AI with a value of 0 was negligible. It indicates that wrong selection by the gate-keeping approach after rejecting H_0 is rare.

Table 2.7. The distribution of selected AI by the gate-keeping method after the Wald test (at 5% level) under balanced randomization and a total sample size of $n = 200$.

	$\Delta = 0.00$		$\Delta = 0.05$						$\Delta = 0.10$					
	DS1-Null		DS1-VP1		DS1-VP2		DS1-VP3		DS1-VP1		DS1-VP2		DS1-VP3	
AI	Value	Prob.	Value	Prob.	Value	Prob.	Value	Prob.	Value	Prob.	Value	Prob.	Value	Prob.
(0;0,0)	0.00	0.006	0.00	0.000	0.00	0.000	0.00	0.001	0.00	0.000	0.00	0.000	0.00	0.000
(0;0,1)	0.00	0.006	0.00	0.000	0.87	0.000	0.93	0.005	0.00	0.000	1.23	0.000	1.32	0.001
(0;1,0)	0.00	0.006	0.00	0.000	1.75	0.001	2.49	0.045	0.00	0.000	2.47	0.000	3.52	0.033
(0;1,1)	0.00	0.006	0.00	0.000	2.62	0.007	3.42	0.209	0.00	0.000	3.70	0.001	4.84	0.300
(1;0,0)	0.00	0.006	4.48	0.169	3.63	0.018	-1.24	0.000	6.33	0.241	5.13	0.009	-1.76	0.000
(1;0,1)	0.00	0.006	4.48	0.168	4.50	0.068	0.31	0.000	6.33	0.251	6.36	0.053	0.44	0.000
(1;1,0)	0.00	0.006	4.48	0.166	5.38	0.119	2.48	0.027	6.33	0.231	7.60	0.120	3.52	0.014
(1;1,1)	0.00	0.006	4.48	0.167	6.25	0.458	4.04	0.385	6.33	0.228	8.83	0.759	5.72	0.598
	$\Delta = 0.00$		$\Delta = 0.05$						$\Delta = 0.10$					
	DS2-Null		DS2-VP1		DS2-VP2		DS2-VP3		DS2-VP1		DS2-VP2		DS2-VP3	
AI	Value	Prob.	Value	Prob.	Value	Prob.	Value	Prob.	Value	Prob.	Value	Prob.	Value	Prob.
(0;0,1)	0.00	0.012	0.00	0.000	0.00	0.000	0.00	0.000	0.00	0.000	0.00	0.000	0.00	0.000
(0;1,1)	0.00	0.012	0.00	0.000	1.92	0.003	3.21	0.181	0.00	0.000	2.77	0.000	4.59	0.170
(1;0,1)	0.00	0.013	4.48	0.380	4.00	0.076	0.40	0.000	6.33	0.488	5.90	0.042	0.57	0.000
(1;1,1)	0.00	0.013	4.48	0.381	5.92	0.678	4.42	0.576	6.33	0.489	8.67	0.932	6.31	0.806
	$\Delta = 0.00$		$\Delta = 0.05$						$\Delta = 0.10$					
	DS3-Null		DS3-VP1		DS3-VP2		DS3-VP3		DS3-VP1		DS3-VP2		DS3-VP3	
AI	Value	Prob.	Value	Prob.	Value	Prob.	Value	Prob.	Value	Prob.	Value	Prob.	Value	Prob.
(0;0,1)	0.00	0.0016	0.00	0.000	0.00	0.000	0.00	0.057	0.00	0.000	0.00	0.000	0.00	0.023
(0;1,1)	0.00	0.0016	0.00	0.001	2.59	0.034	-2.96	0.000	0.00	0.000	3.65	0.012	-4.24	0.000
(1;1,1)	0.00	0.0019	4.48	0.807	5.17	0.776	2.22	0.761	6.33	0.985	7.30	0.972	3.18	0.962

2.4 Application: Selecting the Best Web Design for Smoking Cessation

To demonstrate the use of proposed Wald test in a real data set, we show the application using a data set from a two-stage web-based smoking cessation study conducted by Center for Health Communications Research (CHCR) at University of Michigan (Strecher et al., 2008). The first stage randomly directed subjects to 4 types of website with different designs defined by two two-level behavioral factors, which are personalization of the source materials (high source vs. low source) and the depth of success story narration (high depth vs. low depth). The goal of the first stage was to help smokers to quit in a 6-month window.

- **Personalization of source.** Highly personalized source version website included a

photograph of, and supportive text from, the smoking-cessation team of the HMO, with friendly words like "we" and "our team". Also, it ended the webpage with a signature from the team. While low-personalized version only included a photograph with the build of HMO institution and used impersonal terms, such as "this organization", with no signature.

- **Depth of success story.** Patients were directed to a webpage with a hypothetic story about an individual who successfully quit smoking. A high-depth story referred to a success story tailored to a subject's name, gender, age, ethnicity, martial status, and a series of detailed information. While low-depth stories were only tailored to the subject's name and gender.

After the intermediate evaluation at 6 months after baseline, consenting patients in the study were further randomized to the second stage between a website that included in-depth materials (treatment) and one that contained only a brief message (control), and were followed for another 6 months.

- **Treatment website.** Treatment websites included links to access the materials selected from 8 original Forever Free relapse prevention booklets (Brandon et al. 2012). subjects logged in treatment websites received a welcome page that encouraged them to read any of the 8 web booklets that they felt important in helping in their cessation efforts.

- **Control website.** Control websites only contained a brief message about the subject's current smoking status.

At the end of the second stage, the final smoking status was defined by a 7-day point

prevalence abstinence and dichotomized as success versus failure [8]. Thus, the AI value in our analysis was the probability of quitting smoke successfully in 12 months.

There were 16 AIs in total defined in this two-stage SMART (cf. Table 2.1). The MLEs of all the AI values $\hat{\boldsymbol{\Theta}}$ were calculated using the study data in 282 patients and listed in Table 2.1. The AI with Web-design combining "Low source" and "low depth" at Stage 1 following by a website only contained a brief message about the patient's current smoking status (control) at Stage 2 has the highest estimated value ($\theta_1 = 0.43$). The proposed Wald test was applied and the test statistics was obtained as $Q = 19.14$. As the null distribution of the Wald test is chi-squared with 11 degrees of freedom according to Theorem 2, I obtained $P = 0.059$. Hence, the procedure failed to reject null (5) at 5% significance, and might not proceed to select an intervention. I note that the study was not originally powered for this specific inferential purpose, but rather was analyzed separately by stages (Strecher et al. 2008; Chakraborty, Strecher and Murphy, 2010). Having said that, the omnibus test was quite close to reach significance at 5%. I also compared the AI's in pairwise fashion. The test results against the observed best AI ($g = 1$) are reported in the last column of Table 2.1: Although there was one comparison had a P value less than 0.05 (e.g. the AI with $g = 8$ and $P = 0.044$), multiplicity adjustment would require a P value less then 0.0004 according to the Bonferroni's method.

2.5 Discussion

I have proposed in this chapter a Wald test that can be applied as a gate-keeping test for AI selection in a SMART. Also, I studied the selection properties of a gate-keeping approach based on the proposed Wald test under SMART with varying design structures,

randomization scheme and outcome scenario. As a results of the selection paradigm, one can substantially reduce the sample size of a SMART by powering the study based on a gate-keeping test, and thus improving feasibility of doing a SMART is promising. Simulation study shows that the power of the proposed Wald test is affected by the number of embedded AIs to a lesser extent than pairwise comparison with multiplicity adjustments. As the "curse of dimensionality" is a major concern in evaluating AIs embedded in a SMART, especially if we consider more than two stages and multiple response categories, performing such an omnibus test as a gate-keeping test is a reasonable approach on ground of feasibility.

From a practical viewpoint, the proposed test facilitates clear clinical decisions at the end of a trial. In this article, I consider an approach whereby an AI is selected upon rejecting the null of no difference. I note that the goal of a selection trial is not to select the best intervention with high probability, but rather select an intervention that is not "bad": The two objectives coincide in scenarios where no AI falls in the indifference zone (cf. VP1); see also the examples in Bechhofer, Santner, and Goldsman (1995) and Cheung (2007). The proposed Wald test can be coupled with other clinical decision rules such as identifying inferior interventions, as long as these rules are pre-specified. As a case in point, the re-analysis of smoking cessation (cf. Table 2.1) shows that several AIs ($g = 2, 5$, and 15) had estimated values close to the observed best ($g = 1$), whereas some were clearly inferior to these promising AIs ($g = 8$). Retrospectively, it might be appropriate to perform a unadjusted pairwise tests between the observed inferior AI with the observed best AI, when the null hypothesis of the gate-keeping test in (5) was rejected, and thus identify where the difference lied; e.g., the AI with $g = 8$ would have been declared inferior according to this procedure. This multiple comparison procedure

protect against false positive finding using the proposed Wald test in a similar manner to Fisher's least significant difference approach using ANOVA (Meier, 2006).

Chapter 3 Multiple Comparison with the Best Simultaneous Confidence Intervals to Identify Inferior Adaptive Interventions

3.1 Introduction

I introduced in previous chapter a Wald test built on Maximum likelihood estimator (MLE) of adaptive intervention (AI). The proposed test can be applied as a gate-keeping test for selecting the best intervention in SMART. In an early phase trial with the objective to select one promising or several near-best AIs, the gate-keeping test can help to protect against the inflation of type I error rate due to an exhaustive search. In this chapter, I view the multiple comparison problem from a different angle and propose a new method to address it. Unlike the gate-keeping test that aims to select the best adaptive intervention in SMART and move it forward to the final confirmatory trial, the method proposed in this chapter help to identify the inferior AIs efficiently and then eliminate the inferior AIs from moving to next phase investigation.

Specifically, I developed a method to build simultaneous confidence intervals, in which I adopted the concept of *Multiple Comparisons with the Best (MCB)* simultaneous confidence intervals proposed in Hsu (1984). The MCB concept was derived from the ranking and selection literatures which lends itself to the problem of subset selection in Hsu (1981). The idea is that each intervention is compared to the truly best intervention, which is assumed to be unknown, using confidence interval; any intervention with a confidence interval excluding zero will be identified as inferior to the truely best. The concept of MCB is appealing for a SMART where a clinical trialist expects some AIs would have lower values comparing to the majority of the AIs and thus the goal of the study is to

effectively identify the inferior AIs. While the original method for constructing MCB confidence intervals has been well studied for non-adaptive interventions comparisons, it was built on traditional randomized clinical trial design that randomly assigns patients to several independent groups for comparing non-adaptive interventions. The methods was derived based on known correlation structures. Considering the correlations structure between the estimates of adaptive intervention values derived based on SMART data is typically unknown priori, the original method can not be directly applied to SMART settings. In this chapter, a stepwise algorithm that extends the idea in Edwards and Hsu (1983) to construct MCB intervals for AIs embedded in a SMART setting is proposed.

The method proposed in this chapter is motivated by a clinical examples that extracts data from the CODIACS (Comparison of Depression Intervention after Acute Coronary Syndromes) trial, which provides data that allow to evaluate up to eight adaptive interventions of depression management for patients who had surgeries after acute coronary syndrome (ACS) [11]. Each patient in CODIACS was given medication or problem-solving therapy at baseline and then was re-assigned another treatment based on the response to initial treatment at around 8 weeks after baseline. The objective of the trial is to maximize the reduction in depression measured by change in the Beck Depression Inventory (BDI). The AIs embedded in the CODIACS trial are tabulated in Tables 3.1.

The rest of this chapter is organized as follows. Chapter 3.2 introduces the method of constructing MCB simultaneous confidence intervals. Chapter 3.3 evaluates finite sample performance of the proposed method using simulation. Chapter 3.4 shows an application using the depression management trial data. This chapter ends up with some discussion in Chapter 3.5.

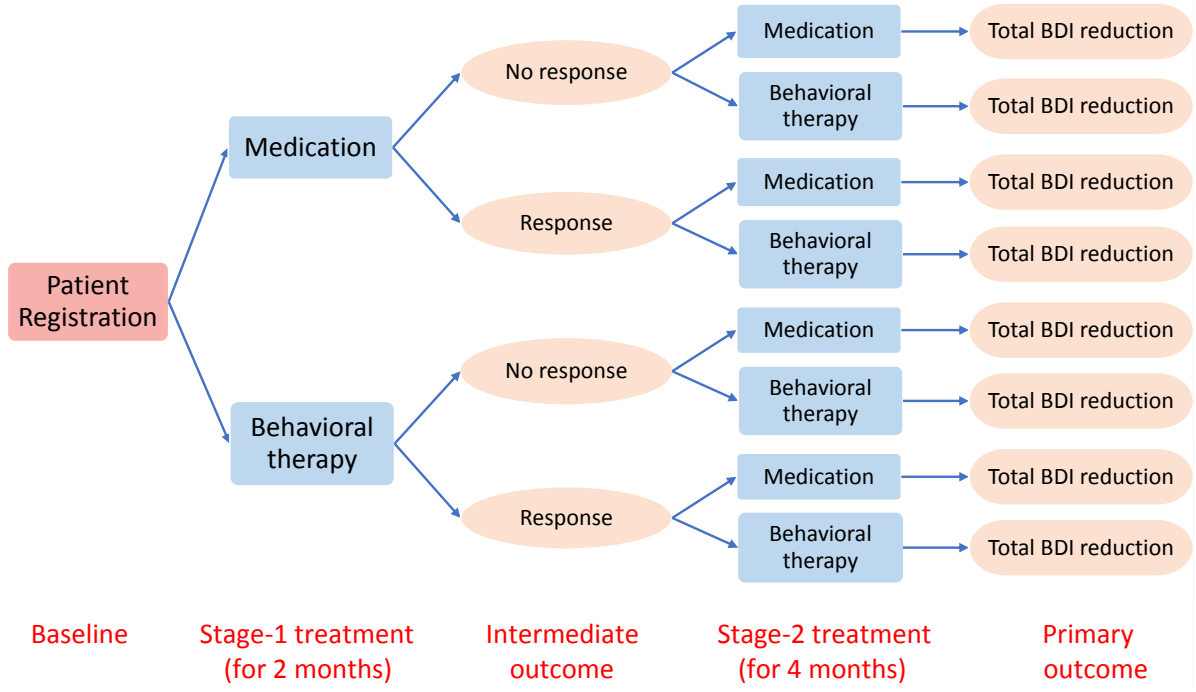


Figure 3.1. Design structures of the CODIACS trial.

Table 3.1. Multiple comparisons of AIs embedded in CODIACS. The MCB intervals for AI g compared it with the truly best AI (assuming unknown), whereas the Bonferroni's intervals compared each AI with the observed best ($g = 5$).

	Stage-1	Stage-2 Treatment for				80% confidence intervals	
AI (g)	Treatment	non-response	response	$\hat{\theta}_g$ (sd)	δ_g	MCB	Bonferroni
1	MED	MED	MED	6.3 (1.1)	1.98	[-19.7, 0.0]	[-25.7, 7.3]
2	MED	MED	PST	3.3 (1.2)	1.99	[-22.7,-0.3]	[-28.7, 4.5]
3	MED	PST	MED	10.7 (0.6)	2.04	[-15.2, 0.0]	[-21.1,11.8]
4	MED	PST	PST	7.8 (1.1)	1.98	[-18.2, 0.0]	[-24.2, 8.8]
5	PST	MED	MED	15.5 (6.0)	1.71	[-7.6, 0.0]	-
6	PST	MED	PST	9.5 (1.0)	2.00	[-16.3, 0.0]	[-22.2,10.2]
7	PST	PST	MED	14.2 (6.1)	1.71	[-8.9, 0.0]	[-3.8, 1.4]
8	PST	PST	PST	8.2 (1.1)	1.98	[-17.6, 0.0]	[-23.6, 9.1]

sd: estimated asymptotic standard deviation of $\hat{\theta}_g$; MED: medication; PST: problem-solving therapy.

3.2 MCB Simultaneous Confidence Intervals

The proposed method of constructing MCB simultaneous confidence intervals was derived based on the same setting, notation and model described in Chapter 2.2.1. Let

$$\boldsymbol{\Theta} = (\theta_1, \ldots, \theta_G), \text{ where } g = 1, \ldots, G,$$

be the vector of the values of all the adaptive interventions embedded in a general two-stage SMART design as shown in Figure 2.1, arranged in a lexicographical order of $(i; k_{i1}, \cdots, k_{iJ_i})$, where g is the AI indicator with value from 1 to G. G is the total number of AIs embedded in a SMART. For each value of g, let $\boldsymbol{D}_g$ be a $(G-1) \times G$ contrast matrix such that its g-th column is a $(G-1)$ vector of 1's, and its j-th column is a $(G-1)$-vector whose j-th entry is -1 and other entries are 0's if $j \leq g-1$; whose $(j-1)$-th entry is -1 and other entries are 0's if $j \geq g+1$. Then

$$\boldsymbol{D}_g \cdot \sqrt{n}(\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}) \xrightarrow{d} N\Big(\mathbf{0},\ \boldsymbol{D}_g \boldsymbol{\Sigma} \boldsymbol{D}_g^T\Big).$$

Let σ_{ig}^2 be the ith diagonal entry of $\boldsymbol{D}_g \boldsymbol{\Sigma} \boldsymbol{D}_g^T$. Write

$$\boldsymbol{D}_g \boldsymbol{\Sigma} \boldsymbol{D}_g^T = \text{diag}\{\sigma_{ig}\} \cdot \boldsymbol{R}_g \cdot \text{diag}\{\sigma_{ig}\},$$

which means that $\boldsymbol{R}_g$ is the correlation coefficient matrix of $N\Big(\mathbf{0},\ \boldsymbol{D}_g \boldsymbol{\Sigma} \boldsymbol{D}_g^T\Big)$. Let $\delta_g > 0$ be the unique solution to the following equation

$$P\Big(\big|Z_{ig}\big| \leq \delta_g;\ i = 1, \ldots, G-1\Big) = 1 - \alpha, \tag{11}$$

where α is the familywise type I error rate and

$$(Z_{1g}, \ldots, Z_{G-1,g})^T \sim N\left(\mathbf{0},\ \hat{\boldsymbol{R}}_g\right),$$

where $\hat{\boldsymbol{R}}_g$ is a consistent estimator of $\boldsymbol{R}_g$. Here δ_g is the $(1-\alpha)$th quantile of a joint normal distribution with dimension $(G-1)$.

The method to construct MCB simultaneous confidence intervals is proposed as following 2 steps:

Step 1: select a subset of AIs, $\mathcal{B}$, defined as

$$\mathcal{B} = \left\{ g : \hat{\theta}_g - \hat{\theta}_i + \delta_g \hat{\sigma}_{ig}/\sqrt{n} > 0;\ g = 1, \cdots, G;\ i = 1, \cdots, G;\ i \neq g \right\}.$$

For the gth AI embedded in a SMART, I construct $(G-1)$ confidence intervals of $(\theta_g - \theta_i)$'s, where $i = 1, \cdots, G$ and $i \neq g$, based on the δ_g obtained from equation (11). If all these $(G-1)$ confidence intervals contain upper limits greater than zero, I select g for $\mathcal{B}$. Otherwise, g is excluded from $\mathcal{B}$. Note that $\mathcal{B}$ may include more than one AI and can not be empty by definition. In a case that an AI has lower limits greater than 0 when it compares with all the other AIs embedded in SMART, the subset $\mathcal{B}$ is singleton and contains only the AI itself. The subset $\mathcal{B}$ can be viewed as a conservative estimates of the truly best AI.

Step 2: given $\mathcal{B}$, the MCB intervals for the gth AI can be obtained by

$$[L_g,\ U_g] = \left[\min_{b \in \mathcal{B}} L_{gb},\ \max_{b \in \mathcal{B}} U_{gb}\right],$$

where

$$L_{gb} = \begin{cases} 0, \text{ if } g = b, \\ (\hat{\theta}_g - \hat{\theta}_b) - \delta_g \hat{\sigma}_{gb}/\sqrt{n}, \text{ if } g \neq b, \end{cases}$$

$$U_{gb} = \begin{cases} 0, \text{ if } g = b, \\ \min\left\{0, \ (\hat{\theta}_g - \hat{\theta}_b) + \delta_g \hat{\sigma}_{gb}/\sqrt{n}\right\}, \text{ if } g \neq b, \end{cases}$$

for $g = 1, \ldots, G$ and $b \in \mathcal{B}$.

By construction, no upper limit of MCB intervals U_g can be positive. This is because that a MCB interval gives an estimated rage of the value difference between an AI and the true optimal AI that is assumed to be unknown.

Theorem 3. As $n \to \infty$, $[L_g, U_g]$, $g = 1, \ldots, G$ is a set of $100(1-\alpha)\%$ asymptotic simultaneous confidence intervals for $\theta_g - \max\limits_{1 \leq j \leq G} \theta_j$, $j = 1, \ldots, G$.

The proof of Theorem 3 is given in Appendix 1. An AI with a negative upper limit U_g, as opposed to having $U_g = 0$, can be concluded as inferior (to the truly best) with confidence. Furthermore, if the subset $\mathcal{B}$ contains only one AI, the MCB interval associated with this AI must be $[0, 0]$, that is, $L_g = 0$.

3.3 Finite Sample Performances

3.3.1 SMART Designs and Outcome Scenarios

Simulation study was conducted to evaluated the finite sample performances of the MCB confidence intervals. Three two-stage design structures were considered in simulation as

shown in Figure 3.2. The first design structure (DS1) mimics the situation described in CODIACS trial, in which there are two treatment options at each decision making point and the intermediate outcomes for both Stage-1 treatment are defined as binary variable. That is to say, $T_i, R_{ij}, S_{ijk} \in \{0, 1\}$ for $i, j, k = 1, 2$. DS2 is similar to DS1 but respondents will keep using the same treatment as received in Stage 1. Under DS3, patients who are randomized to received T_1 will be classified as responders and non-responders. Non-responders will be further randomized to receive 2 treatments at Stage 2, while responders will receive the same treatment as Stage-1 treatment. For those patients who are randomly assigned to receive T_2, they will be given exactly the same treatment at Stage 2. DS1, DS2 and DS3 provide data that allow to evaluate 8, 4 and 3 AIs, respectively.

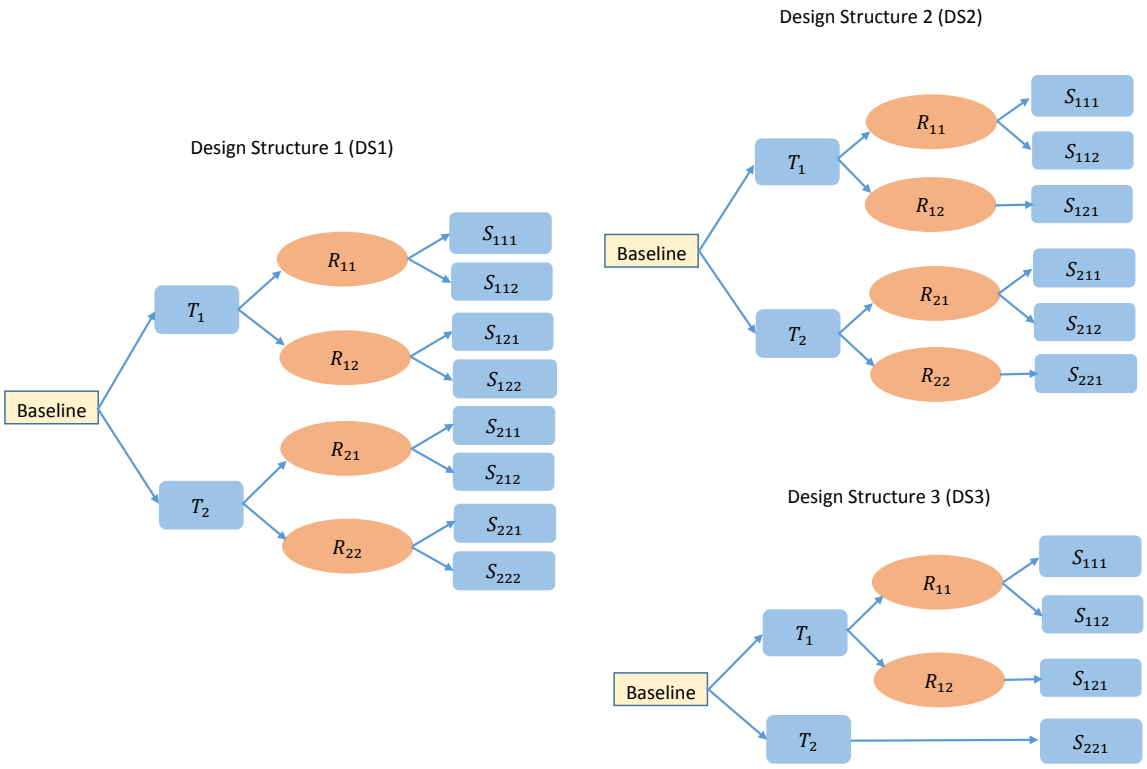


Figure 3.2. Design structures considered in the simulation.

Three sets of randomization schemes were considered for each design structure. First, we considered a balanced randomization (BR) scheme, under which the randomization probabilities of both treatment coded as 0 and 1 were set to be 0.5. That is to say, at each randomization point, the chances that a patient was randomly assigned to receive treatment coded as 0 and 1 were equal. Second, we considered an unbalanced randomization (UBR) scheme, where the randomization probabilities corresponding to the treatment options coded as 0 were set to be 0.3 and those corresponding to the treatment options coded as 1 were set to be 0.7. Third, we considered $P(V = 0) = P(V = 1) = 0.5$ at Stage 1, while $P(V = U|U, X = 0) = 0.3$ and $P(V = U|U, X = 1) = 0.7$ at Stage 2. Under this randomization scheme, the probability that a responder will receive the same treatment at Stage 2 as Stage-1 treatment will increase to 0.7 and the probability that a non-responder will received the same treatment at Stage 2 as Stage-1 treatment will decrease to 0.3, and thus Stage-1 implements a randomized play-the-winner (RPTW) rule when the first and the second treatment options are identical. Therefore, the three designs (DS1, DS2 and DS3) and the three randomization schemes (BR, UBR and RPTW) yielded 9 SMART designs in simulation.

Figure 3.3 gives the 3 outcome scenarios considered for each SMART design. Under Value Pattern 1 (VP1; top panel), AIs with the same stage-1 treatment had the same values and AIs started with treatment coded as 1 had greater values comparing to those started with 0; under VP2, the values of the AIs were uniformly higher if their stage-1 treatment was $U = 1$; under VP3 (bottom panel), the best AI had stage-1 treatment $U = 1$ while the second best had stage-1 treatment $U = 0$, and so on and so forth, following an alternating pattern. For each value pattern I considered effect sizes $\Delta = 0.05$ and 0.10. More details of outcome scenarios see Chapter 2.3.2 and Appendix 2.

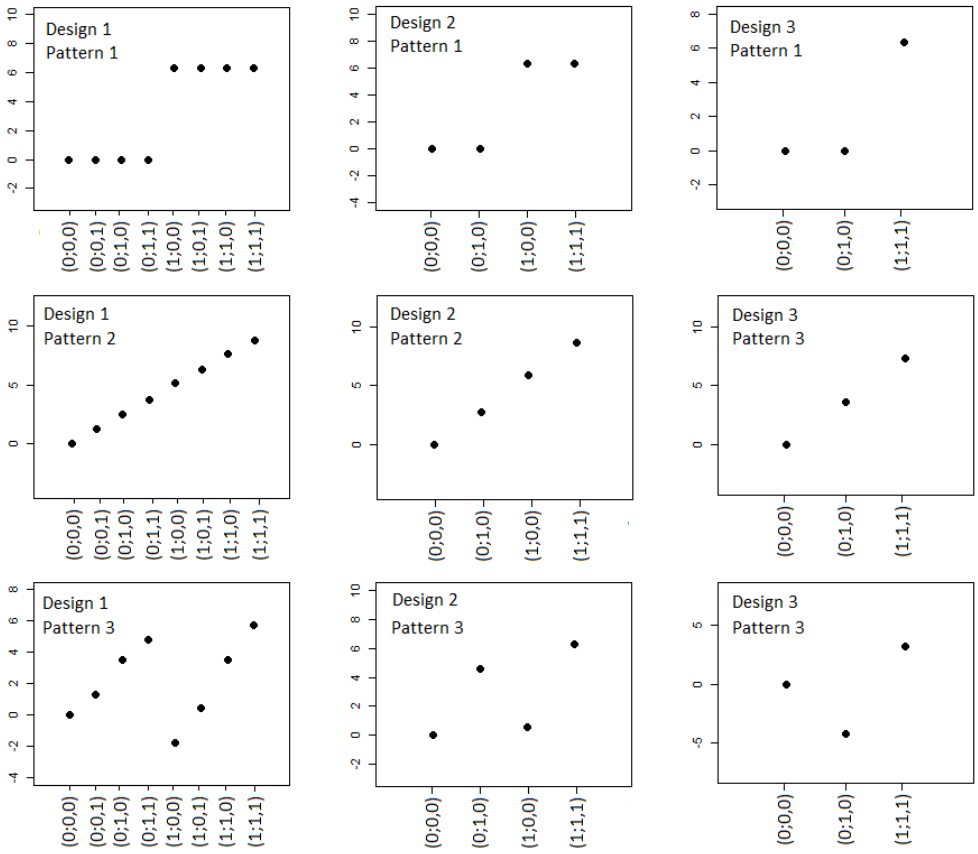


Figure 3.3. Value patterns considered in the simulation.

3.3.2 Simulation Results

I evaluated the finite sample performances of the proposed method to constructure MCB confidence intervals for identifying inferior AIs: AIs with MCB confidence intervals excluding zero will be declared as inferior to the truly best AI and removed from further considerations. To anticipate the clinical context where there are many candidates of the best AI and the goal is to move forward to the next phase investigation with a subset, we may consider applying the MCB intervals at a confidence level less than 95% so as to afford a higher differentiating power. Specifically, I considered 80% confidence in this simulation.

Table 3.2. Properties of 80% MCB intervals and simultaneous confidence intervals using Bonferroni's adjustment with $n = 200$: coverage probability (cov) and average width of intervals (wid).

			$\Delta = 0.05$				$\Delta = 0.10$			
Design	Value	Randomization	MCB		Bonferroni		MCB		Bonferroni	
structure	pattern	scheme	cov	wid	cov	wid	cov	wid	cov	wid
DS1	VP1	BR	0.927	6.63	0.918	5.42	0.920	6.95	0.918	5.42
		UBR	0.941	7.07	0.927	6.31	0.925	7.47	0.927	6.31
		RPTW	0.929	7.13	0.920	5.72	0.921	7.52	0.920	5.72
DS1	VP2	BR	0.901	6.64	0.915	8.63	0.867	7.10	0.915	8.65
		UBR	0.913	7.24	0.924	9.52	0.873	7.73	0.921	9.53
		RPTW	0.907	7.13	0.917	9.57	0.876	7.67	0.916	9.59
DS1	VP3	BR	0.936	6.64	0.905	8.81	0.903	7.30	0.903	8.99
		UBR	0.960	7.70	0.919	9.63	0.946	8.37	0.915	9.75
		RPTW	0.942	7.10	0.904	9.74	0.914	7.77	0.901	9.90
DS2	VP1	BR	0.881	4.95	0.866	3.90	0.880	5.09	0.866	3.90
		UBR	0.885	5.36	0.864	4.80	0.870	5.54	0.864	4.80
		RPTW	0.883	5.28	0.858	3.82	0.877	5.44	0.858	3.82
DS2	VP2	BR	0.858	5.00	0.889	5.75	0.848	5.19	0.893	5.88
		UBR	0.865	5.30	0.878	6.02	0.846	5.48	0.879	6.11
		RPTW	0.838	5.37	0.877	6.47	0.835	5.65	0.882	6.60
DS2	VP3	BR	0.872	4.97	0.873	5.67	0.857	5.23	0.876	5.72
		UBR	0.886	5.35	0.876	5.98	0.867	5.67	0.880	6.03
		RPTW	0.856	5.30	0.869	6.41	0.840	5.61	0.872	6.45
DS3	VP1	BR	0.806	3.52	0.843	4.00	0.805	3.468	0.841	3.999
		UBR	0.813	4.31	0.852	4.79	0.808	4.232	0.852	4.786
		RPTW	0.807	3.70	0.843	4.17	0.805	3.643	0.843	4.172
DS3	VP2	BR	0.821	3.66	0.846	4.01	0.815	3.616	0.846	4.022
		UBR	0.839	4.55	0.854	4.80	0.828	4.582	0.854	4.815
		RPTW	0.821	3.75	0.848	4.19	0.818	3.735	0.850	4.197
DS3	VP3	BR	0.809	3.67	0.841	4.02	0.798	3.664	0.840	4.033
		UBR	0.815	4.33	0.845	4.81	0.801	4.390	0.844	4.827
		RPTW	0.821	3.94	0.833	4.19	0.800	3.928	0.829	4.209

Table 3.2 gives the coverage probabilities of the 80% MCB intervals under the same designs and outcome scenarios as in Chapter 3.3.1. To be more concise, each probability pertained to simultaneous coverage, and was calculated as the proportion of simulated trials under which the gth MCB interval covers the corresponding true values of $\theta_g - \max_{1 \le i \le G} \theta_i$ for *all* $g = 1, \ldots, G$. Recall that $G = 8, 4, 3$ under DS1, DS2, and DS3, respectively. For comparison purposes, I also considered simultaneous confidence intervals based on Bonferroni's correction: for each pair of AIs, a confidence interval for the difference of their values was evaluated with confidence level $100\,[1 - 0.2/\{G(G-1)/2\}]\,\%$

so that the overall nominal coverage is 80%. The coverage probability of the Bonferroni intervals was calculated as the proportion of simulated trials under which all $G(G-1)/2$ intervals covered the corresponding true differences. While I note that the Bonferroni's simultaneous confidence intervals address a different estimation problem than the MCB intervals, both methods are valid in that their corresponding coverage probabilities were at least 80% in all scenarios. Indeed, both methods appeared to be conservative, especially under DS1 where there were many AIs. For MCB confidence intervals, the conservativeness was due to the asymptotic approximation: simulation with larger sample sizes showed that the coverage probability approached the nominal 80% as the total sample size n increased.

I also calculated the average widths of the confidence intervals as a measure of efficiency. For the MCB intervals, the average width was taken over the G MCB intervals. For the Bonferroni's pairwise intervals, the average width was taken over the G intervals that compared to the *observed* best AI. I note that this comparison was unfair against MCB intervals in two ways due to the interpretations of these intervals. First, the MCB method did not assume the knowledge of the *true* best AI, and the calculation of MCB intervals implicitly accounted for variability induced due to this unknown parameter. In contrast, the Bonferroni's method avoided estimating the unknown true best AI by comparing all AI's against a known and observed best, and thus was addressing an easier inferential problem. Second, by definition of the Bonferroni pairwise intervals, the observed best when compared to itself would have a width of zero, which would artificially shrink the average width towards a smaller value. However, I opted to use the Bonferroni's intervals as a benchmark to evaluate how the MCB intervals perform under different scenarios. Having noted the difference in interpretations of the two interval es-

timation procedures and the average widths, I observed that the 80% MCB intervals had smaller average width than that of the G corresponding Bonferroni's intervals under value patterns VP2 and VP3, but had larger average width under VP1. As discussed above, since the MCB intervals implicitly account for the variability in estimating what the true best AI is, I would expect larger variability when the true best AI is not unique and thus difficult to estimate, i.e., under VP1. In practice, scenarios such as VP1 where many AIs have the same value are conceivably less likely than the other patterns, especially when the AIs consist of components of different treatment types, such as pharmacological versus behavioral.

Table 3.3 presents the probability of an AI being declared as inferior to the truly optimal AI using 80% MCB intervals under balanced randomization with $n = 200$. Under the null hypothesis ($\Delta = 0$), where all AIs had the same value, the probabilities of erroneously declaring an AI as inferior were less than 5% under all design structures. As the effect size Δ increased, the truly inferior AIs were correctly identified with increasing probabilities. Specifically, under $\Delta = 0.10$ with $n = 200$, the MCB method identified the true worst AI as inferior with probabilities between 73% and 98%.

Table 3.3. The probability of an AI being declared inferior using 80% MCB intervals under balanced randomization and a total sample size of $n = 200$.

	$\Delta = 0.00$		$\Delta = 0.05$						$\Delta = 0.10$					
	DS1-Null		DS1-VP1		DS1-VP2		DS1-VP3		DS1-P1		DS1-VP2		DS1-VP3	
AI	Value	Prob.	Value	Prob.	Value	Prob.	Value	Prob.	Value	Prob.	Value	Prob.	Value	Prob.
(0;0,0)	0.00	0.005	0.00	0.382	0.00	0.461	0.00	0.197	0.00	0.781	0.00	0.857	0.00	0.460
(0;0,1)	0.00	0.006	0.00	0.391	0.87	0.324	0.93	0.098	0.00	0.785	1.23	0.709	1.32	0.236
(0;1,0)	0.00	0.005	0.00	0.379	1.74	0.227	2.49	0.038	0.00	0.778	2.47	0.533	3.52	0.080
(0;1,1)	0.00	0.005	0.00	0.381	2.62	0.156	3.42	0.025	0.00	0.785	3.70	0.361	4.84	0.046
(1;0,0)	0.00	0.006	4.48	0.014	3.63	0.122	-1.24	0.343	6.33	0.015	5.13	0.299	-1.76	0.726
(1;0,1)	0.00	0.006	4.48	0.013	4.50	0.050	0.31	0.115	6.33	0.015	6.36	0.111	0.44	0.334
(1;1,0)	0.00	0.006	4.48	0.015	5.38	0.015	2.49	0.034	6.33	0.017	7.60	0.046	3.52	0.074
(1;1,1)	0.00	0.007	4.48	0.017	6.25	0.001	4.04	0.005	6.33	0.019	8.83	0.000	5.72	0.005
	$\Delta = 0.00$		$\Delta = 0.05$						$\Delta = 0.10$					
	DS2-Null		DS2-VP1		DS2-VP2		DS2-VP3		DS2-VP1		DS2-VP2		DS2-VP3	
AI	Value	Prob.	Value	Prob.	Value	Prob.	Value	Prob.	Value	Prob.	Value	Prob.	Value	Prob.
(0;0,1)	0.00	0.023	0.00	0.667	0.00	0.734	0.00	0.549	0.00	0.943	0.00	0.976	0.00	0.885
(0;1,1)	0.00	0.023	0.00	0.663	1.92	0.440	3.21	0.119	0.00	0.945	2.77	0.759	4.58	0.191
(1;0,1)	0.00	0.021	4.48	0.033	4.00	0.234	0.40	0.442	6.33	0.034	5.90	0.462	0.57	0.773
(1;1,1)	0.00	0.022	4.48	0.041	5.92	0.001	4.42	0.008	6.33	0.041	8.67	0.000	6.30	0.003
	$\Delta = 0.00$		$\Delta = 0.05$						$\Delta = 0.10$					
	DS3-Null		DS3-VP1		DS3-VP2		DS3-VP3		DS3-VP1		DS3-VP2		DS3-VP3	
AI	Value	Prob.	Value	Prob.	Value	Prob.	Value	Prob.	Value	Prob.	Value	Prob.	Value	Prob.
(0;0,1)	0.00	0.044	0.00	0.830	0.00	0.820	0.00	0.401	0.00	0.987	0.00	0.983	0.00	0.630
(0;1,1)	0.00	0.045	0.00	0.835	2.59	0.492	-2.97	0.830	0.00	0.987	3.65	0.748	-4.24	0.982
(1;1,1)	0.00	0.042	4.48	0.000	5.17	0.000	2.23	0.001	6.33	0.000	7.30	0.000	3.18	0.000

3.4 Application: Identifying Inferior AIs for Depression Management

Cheung et al. (2015) analyzed data in a subset of patients enrolled to the CODIACS trial with an objective to further determine which stepped care depression management regimens should be used and which should be discontinued in an implementation stage. A specific task may thus be formulated as eliminating inferior AI's from further practice based on reduction of Beck Depression Inventory at 6 months, which was the primary endpoint in the original study. The value of an intervention in this application is the expected reduction of the depression score. Higher BDI reduction indicates better effect of depression management, and thus the AI leading to the less BDI reduction is viewed as inferior to the optimal AI in this case. Furthermore, each AI would adapt to an initial

response at 8 weeks defined as no increase in depression.

Table 3.1 shows the 80% MCB intervals for the eight possible two-stage AIs embedded in the study. This analysis identified an inferior AI, namely, the AI with $g = 2$ that would start with medication, stay with it upon a non-response, and switch to problem-solving therapy upon a response. I emphasize that this analysis was not intended to estimate the true best AI with statistical confidence. Rather, the utility of this analysis is to exclude the inferior AI from further practice from a quality assurance viewpoint.

As a comparison, I applied the pairwise confidence intervals with Bonferroni's correction described in Chapter 3.3.2. Table 3.1 also gives the intervals that compare each AI with the *observed* best ($g = 5$), and shows that the Bonferroni's correction failed to differentiate any AIs. From an estimation viewpoint, the MCB intervals give more precise estimates than Bonferroni's: the average widths of the 8 MCB intervals was 15.7, and the average width of the corresponding Bonferroni's intervals was 25.3, assuming the width corresponding to $g = 5$ is 0.

Finally, while the goal of the analysis in this case is not to produce a definitive statement about whether the eight AIs were significantly different, I also applied the likelihood-based Wald test proposed in Chapter 2.2 to the CODIACS data for illustration purpose. Based on the estimated AI values in (8) and the asymptotic covariance matrix

$$\hat{\boldsymbol{\Sigma}} = \begin{pmatrix} 1.23 & 0.63 & 0.37 & -0.23 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.63 & 1.54 & 0.01 & 0.91 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.37 & 0.01 & 0.41 & 0.05 & 0.00 & 0.00 & 0.00 & 0.00 \\ -0.23 & 0.91 & 0.05 & 1.19 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 36.42 & 0.58 & 36.23 & 0.39 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.58 & 1.03 & 0.25 & 0.70 \\ 0.00 & 0.00 & 0.00 & 0.00 & 36.23 & 0.25 & 36.95 & 0.97 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.39 & 0.70 & 0.97 & 1.28 \end{pmatrix},$$

I got the Wald test statistic $Q = 36.0$ per Equation (6), compared it against a chi-squared distribution with 5 degrees of freedom according to Theorem 2, and obtained $P < 0.001$. This analysis thus confirmed that overall the values of the AIs were statistically significantly different.

3.5 Discussion

I proposed in this chapter an effective tool that can help to identify the AIs inferior to the truly best AI embedded in a SMART. As one may view a SMART as a study in a series of experiments, identify the inferior AIs and eliminate them from move forward can be an temporary goal of an early phase trial of experimental series. When there are potentially many treatment options, one may want to eliminate inferior AIs often with a slightly lower confidence level than the conventional 95%. The rationale is that the clinical investigation can quickly zero in on the promising interventions. At a later phase, one may aim to perform a selection trial with a goal to move a single AI to the final

confirmatory stage. Specifically, I considered here 80% confidence level in the simulation studies and application.

I have developed and extended the distributional theory of MLE for the value of an AI under general SMART designs in Chapter 2, on which I conducted the inferential procedure of MCB simultaneous confidence intervals. Similar to the method proposed in Chapter 2, a gate-keeping test for selecting the best AI in SMART, the proposed method of constructing MCB intervals can also be generalized to different types of outcomes, such as continuous, binary and count data. Also, the validity of the method proposed in this chapter lies on the full specification of the model, which may be perceived as restrictive in application.

Chapter 4 SRT - An R Package for Implementing SMART

4.1 Introduction

Motivated by providing a user-friendly statistical software to help clinical investigators to design and analyze SMART, an R package, **SRT** (**S**equential Multiple Assignment **R**andomized **T**rial), was developed during my thesis research. In this chapter, I will introduce the usage of this R package using the data from examples in previous two chapters. Functions built in this R package cover 3 major statistical work in SMART, including clinical trial design (i.e., sample size calculation, power calculation), data analysis (i.e., descriptive statistics, global tests, pairwise tests, simultaneous confidence intervals), and data visualization (i.e., design diagram, exploratory data analysis). Most commonly used statistical methods in SMART can be implemented by using this package.

By the time SRT was developed, several R packages related to AI research have been published on the Comprehensive R Archive Network (CRAN). Some packages had been developed for the secondary analysis in observational studies, while others were designed to implement certain analytical methods related to a specific paper. Tang and Melguizo (2015) developed an R package, DTR (**D**ynamic **T**reatment **R**egimes), for implementing the statistical estimating and testing procedures to compare adaptive interventions with survival outcomes. Linn, Laber and Stefanski (2015) created an R package, iqLearn (**I**nteractive **Q**-**Learn**ing), to estimate the values of adaptive interventions based on the Q-learning method, which allowed to incorporate the impacts from multiple covariates in estimation and thus can be applied to secondary analysis. Holloway et al. (2017) developed an R package, named DynTxRegime (**Dyn**amic **T**reatment **Regime**), to evaluate

AI based on observation analysis. A series of methods to estimate the optimal AI in observational studies were built in this package, including Inverse Probability Weighted Estimation, IQ-learning, Q-learning, and some other regression-based methods. Wallace, Moodie and Stephens (2017) developed an R package with the name called DTReg (**D**ynamic **T**reatment **R**egimes Estimation via **Reg**ression-based Techniques), including a series of functions focusing on regression-based estimation and a variance estimation based on bootstrapping.

Comparing to currently existing R packages related to AI research, the SRT package has many innovative features. First, it is innovative to build functions that can help to conduct exploratory data analysis (EDA) for SMART data. There is no R package so far providing graphical tools to present the features of SMART data in this manner. EDA techniques are efficient tools to explore the features of clinical trial data at the beginning of the analysis. SRT allows to output the graphs of descriptive statistics of SMART data from two different angles, by treatment sequences and by AIs. Also, it can output the design diagrams of SMARTs, which can not only show the design structures of SMART data, but also help to design SMARTs. Second, it is innovative to include functions to construct simultaneous confidence intervals for AI comparisons using SMART data. It is common to calculate the confidence intervals of group comparisons adjusted for multiplicity in analyzing RCT data. Unlike the existing packages such that focus on regression-based methods, SRT allows to output 2 types of simultaneous confidence intervals for comparing AI values. Third, it is innovative to include functions that helps to design SMARTs. All the existing R packages focus on data analysis, SRT not only provides functions to display the design structures of SMARTs, but also can help to conduct sample size calculation and power calculation in different fashions, which is

useful at design stage.

The rest of this chapter are organized as follows. Chapter 4.2 illustrates the notations, which are different from the ones I used in previous two chapters for illustration purpose. Chapter 4.3 introduces the format of input data. Chapter 4.4 introduces functions for generating descriptive statistics. Chapter 4.5 illustrates how to compare AIs using functions in SRT, and Chapter 4.6 shows how to use SRT to conduct sample size calculation and power calculation. Considering the goal of this chapter is to introduce the package and to illustrate the usage of it, the theoretical and technical details will be skipped. Readers with interests can find these details in other chapters and appendix.

4.2 Notation

For illustration purpose, the proposed methods in previous two chapters were derived based on a general two-stage SMART designs, under which the stage-1 treatments are non-adaptive and the stage-2 treatments are adapted to the categorized intermediate outcomes. The set of notations

$$\{T_i, R_{ij}, S_{ijk}\}, \text{ where } i = 1, \ldots, I;\ j = 1, \ldots, J_i;\ k = 1, \ldots, K_{ij} \tag{12}$$

were thus used to depict the design structure of a SMART. On the other hand, in clinical trial practice, it is possible to change certain features of a SMART by controlling the values of some design parameters. For example, we could design a two-stage SMART, under which the stage-1 treatment is adapted to the baseline information of individuals, such as age or disease stage of cancer. Also, it is sometime helpful to design a SMART with more than two treatment stages due to the AIs of interest. The package SRT can

handle these variations. To reflect this point, a set of notations that are different from (12) are used in this chapter. I define a stage as an interval of time during which a patient receives one specific treatment. For a T-stage SMART with total sample size n, at the beginning of Stage t, where $t = 1, ..., T$, an evaluation is made for the lth patient and the observed result is denoted by O_{tl}. Let A_{tl} be the treatment given to patient l at Stage t. Also let Y_l be the primary outcome observed on patient l at the end of the study. Data collected from the lth patient who completes a T-stage SMART can be summarized as a longitudinal trajectory as

$$\{O_{1l}, A_{1l}, ..., O_{Tl}, A_{Tl}, Y_l\}, \text{ where } l = 1, \ldots, n. \tag{13}$$

An over-bar above a Latin letter is used here to indicate the history of certain measures up to the current time point, at which the event represented by the Latin letter happens. For example, $\bar{A}_{2l} = \{A_{1l}, A_{2l}\}$ indicates the history of treatments assigned to the lth patient up to Stage 2 and $\bar{O}_{2l} = \{O_{1l}, O_{2l}\}$ represents the results of evaluation for the lth patient at the beginning of Stage 1 and 2. In this fashion, I can simplify (13) as

$$\{\bar{O}_{Tl}, \bar{A}_{Tl}, Y_l\}, \text{ where } l = 1, \ldots, n. \tag{14}$$

I would further suppress the patient indicator l from these notations for convenience when it is appropriate, so that (13) and (14) become $\{O_1, A_1, ..., O_T, A_T, Y\}$ and $\{\bar{O}_T, \bar{A}_T, Y\}$, respectively. Let $d_t(\bar{O}_t, \bar{A}_{t-1})$ be the decision of selecting action $\bar{A}_t$ conditioning on the clinical history of $\{\bar{O}_t, \bar{A}_{t-1}\}$ by Stage t. An AI is a collection of decisions that can be denoted by

$$D = \{d_1(\bar{O}_1); ...; d_T(\bar{O}_T, \bar{A}_{T-1})\}, \tag{15}$$

wherein the stage-specific decision, $d_t(\bar{O}_t, \bar{A}_{t-1})$, can be a scalar or vector, depending on the number of combinations of $(\bar{O}_t, \bar{A}_{t-1})$. For example, an AI in CODIACS can be $\{d_1; d_2(A_1 = d_1, O_2 = 0), d_2(A_1 = d_1, O_2 = 1)\}$, under which the stage-1 decision rule is a scalar and the stage-2 decision rules is a vector of 2 elements. It is common in practice to set $O_t|(\bar{O}_t, \bar{A}_{t-1})$ as binary for the purpose of reducing the dimensionality of SMART data. Also, in some trials (e.g., CODIACS and CHCR), the stage-1 treatment were designed as non-adaptive. Therefore, O_1 does not exist and (13) becomes

$$\{A_{1l}, O_{2l}, A_{2l}, \ldots, O_{Tl}, A_{Tl}, Y_l\}, \text{ where } l = 1, \ldots, n.$$

4.3 Input SMART Data

The functions of package SRT require to input wide format SMART data with variables shown in (16), where O_t is the tailoring variable measured at the beginning of Stage t, A_t is the treatment assignment for Stage t for $t = 1, \ldots, T$, and Y is the final primary outcome. For users who start analyzing a SMART data using SRT, if the original SMART data are long format, it is required to transform the data into wide format such that each row of the data represents one patient. Also, if the variable names in the original data are different from those in (16), it is important to rename the variables as in (16).

$$\{O1, A1, O2, A2, ..., OT, AT, Y\}. \tag{16}$$

In a case such that the stage-1 treatment is not adaptive to baseline information, there is no $O1$ in input data and thus the variables become

$$\{A1, O2, A2, ..., OT, AT, Y\}. \tag{17}$$

For illustration purpose, I generated a pseudo SMART data under the design structure shown in Figure 4.1, wherein the stage-1 treatment A_1 is not adaptive and has two treatment options, coded as $A_1 = 0$ and $A_1 = 1$. The intermediate outcomes O_2 is binary for both $A_1 = 0$ and $A_1 = 1$, where $O_2 = 1$ for response and $O_2 = 0$ for no response. Given each combination of (A_1, O_2), there is always 2 options of stage-2 treatment A_2, coded as $A_2 = 0$ and $A_2 = 1$. Figure 4.2 gives the R programs to generate the pesudo SMART data. There are 200 patients in this data set. Each patient followed a treatment sequence (A_1, O_2, A_2). The primary outcome Y is continuous and drawn from a normal distribution randomly with mean $\mu = (0, 1.23, 2.47, 3.70, 5.13, 6.36, 7.60, 8.83)$ and standard deviation $\sigma = 10$. The bottom of Figure 4.2 gives the first 10 rows of this data, with variables A1, O2, A2 and Y.

Figure 4.1. Design structure of a pseudo two-stage SMART data

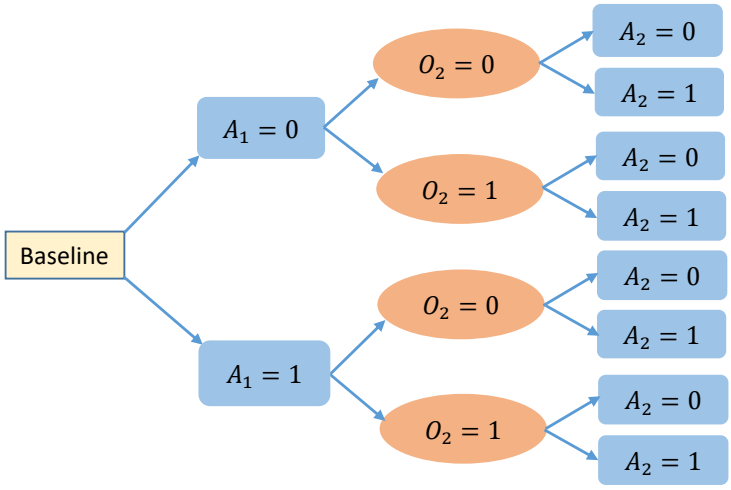


Figure 4.2. Generate a pseudo SMART data

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function | Addins

Source

Console ~/

```
>
> N=200  #total sample size

> A1=O2=A2=Y=rep(NA,N)
> dat=data.frame(A1,O2,A2,Y)  #SMART data

>
> #stage-1 treatment
> pi1=c(1,1)
> dat$A1=sample(c(0,1),size=N,prob=pi1,replace=T)
> N0=length(dat$A1[which(dat$A1==0)]); N1=N-N0
>
> #Intermediate outcome
> dat$O2[which(dat$A1==0)]=rbinom(n=N0,size=1,prob=1/3)
> dat$O2[which(dat$A1==1)]=rbinom(n=N1,size=1,prob=1/3)
>
> N00=nrow(dat[which(dat$A1==0 & dat$O2==0),]); N01=N0-N00
> N10=nrow(dat[which(dat$A1==1 & dat$O2==0),]); N11=N1-N10
>
> #stage-2 treatment
> pi2=c(1,1)
> dat$A2[which(dat$A1==0 & dat$O2==0)]=sample(c(0,1),size=N00,prob=c(1,1),replace=T)
> dat$A2[which(dat$A1==0 & dat$O2==1)]=sample(c(0,1),size=N01,prob=c(1,1),replace=T)
> dat$A2[which(dat$A1==1 & dat$O2==0)]=sample(c(0,1),size=N10,prob=c(1,1),replace=T)
> dat$A2[which(dat$A1==1 & dat$O2==1)]=sample(c(0,1),size=N11,prob=c(1,1),replace=T)
>
> N000=nrow(dat[which(dat$A1==0 & dat$O2==0 & dat$A2==0),]); N001=N00-N000
> N010=nrow(dat[which(dat$A1==0 & dat$O2==1 & dat$A2==0),]); N011=N01-N010
> N100=nrow(dat[which(dat$A1==1 & dat$O2==0 & dat$A2==0),]); N101=N10-N100
> N110=nrow(dat[which(dat$A1==1 & dat$O2==1 & dat$A2==0),]); N111=N11-N110
>
> #primary outcome
> dat$Y[which(dat$A1==0 & dat$O2==0 & dat$A2==0)]=rnorm(N000,mean=0.00,sd=10)
> dat$Y[which(dat$A1==0 & dat$O2==0 & dat$A2==1)]=rnorm(N001,mean=3.37,sd=10)
> dat$Y[which(dat$A1==0 & dat$O2==1 & dat$A2==0)]=rnorm(N010,mean=0.00,sd=10)
> dat$Y[which(dat$A1==0 & dat$O2==1 & dat$A2==1)]=rnorm(N011,mean=3.37,sd=10)
> dat$Y[which(dat$A1==1 & dat$O2==0 & dat$A2==0)]=rnorm(N100,mean=5.13,sd=10)
> dat$Y[which(dat$A1==1 & dat$O2==0 & dat$A2==1)]=rnorm(N101,mean=8.83,sd=10)
> dat$Y[which(dat$A1==1 & dat$O2==1 & dat$A2==0)]=rnorm(N110,mean=5.13,sd=10)
> dat$Y[which(dat$A1==1 & dat$O2==1 & dat$A2==1)]=rnorm(N111,mean=8.83,sd=10)
> dat$Y=round(dat$Y,2)
>
> dat[1:10,]
   A1 O2 A2      Y
1   1  0  1   6.00
2   0  0  0   8.29
3   0  0  1   5.86
4   1  0  0  31.76
5   0  1  1   6.77
6   1  1  0 -12.59
7   1  1  1   7.50
8   0  1  1   6.11
9   1  1  1   7.51
10  0  0  1  10.09
>
```

4.4 Descriptive Statistics

Although the ultimate goal of analyzing SMART data is identifying the best (or inferior) AI based on the statistical inferences of comparing AI values, the first step of analysis in SMART is usually summarizing the features of data by looking at descriptive statistics. SRT provides two functions, **seqmeans(.)** and **atsmeans(.)**, to summarize the sequence-specific and the AI-specific descriptive statistics. Both functions have options to output graphs.

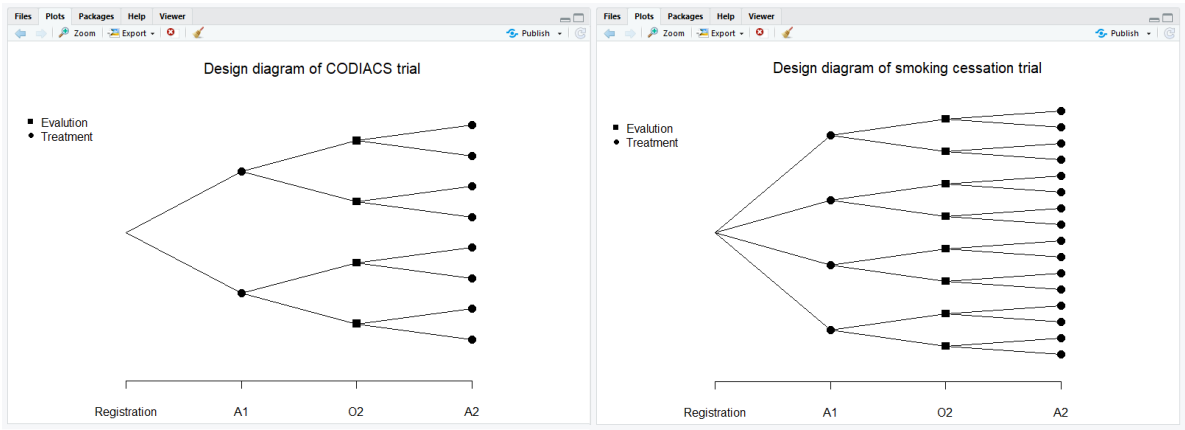
An T-stage AI consists of multiple treatment sequences, each of which is indexed by a series of stage-specific intermediate outcomes and treatments, $(\bar{O}_T, \bar{A}_T)$, occurred at stage $t = 1, \ldots, T$. A patient who completes a SMART will experience one treatment sequence. Specifically, which treatment sequence a patient will follow in a SMART, depends on the results of sequential randomization and the intermediate responses observed on this patient. The function **seqmeans(.)** outputs all the treatment sequences embedded in SMART arranged in the lexicographical order of $(\bar{O}_T, \bar{A}_T)$, and returns the descriptive statistics per each treatment sequence. For each sequence, it displays the values of $(\bar{O}_T, \bar{A}_T)$, the number of patients (N), sample mean (MEAN) and sample variance (VAR) of primary outcome. The option of "family" allows users to specify the type of primary outcome as either continuous (family="normal") or binary (family="binomial"). The default of this option is continuous. This function also contains an option "plot" that allows to choose the output graph. If we choose plot="d", the function will give design diagram of SMART. If choose plot="s", the function will give box plots for continuous primary outcome and bar charts for binary outcome by sequence, depending on the value of option "family". The default of option "plot" is design diagram.

Figure 4.3 shows two examples of outputs of **seqmeans(.)**, using the data of depression management trial (upper panel) and smoking cessation trial (lower panel). The windows on the left-hand side show the R programs and output descriptive statistics in console for two examples, where I chose to output graphs of descriptive statistics and specify family="normal" for the depression management trial and family="binomial" for the smoking cessation trial. Since the primary outcome of depression management trial is continuous, when I set plot="s", it output the box plots by treatment sequence as shown in the upper-right window. Similarly, since the primary outcome is binary in the smoking cessation case, the box plots were replaced by bar charts as shown on the lower right window.

Figure 4.3. Output of function seqmeans(.)

Figure 4.4 are the design diagrams output by using the function **seqmeans(.)** with option plot="d" for both the depression management trial and the smoking cessation trial. The dots represent the stage-specific treatments and squares represent the intermediate outcomes. From the design diagram, we can see that there are two treatment options at all the decision making points and the intermediate outcomes are binary in the depression management trial. While, there are four stage-1 treatment options and the intermediate outcomes are also binary for the smoking cessation trial. In addition, for each decision making point at Stage 2, there are two treatment options in this case. Both the stage-1 treatments in the depression management trial and the smoking cessation trial are not adapted to the baseline information. The lines connecting the stage-specific treatment and intermediate outcome from left to the right correspond to one treatment sequence. It can see that there are totally 8 treatment sequences in the left diagram and 16 treatment sequences in the right diagram.

Figure 4.4. Design diagram by seqmeans(.)



It is important to estimate the values of AIs and observe the pattern of these estimates. The R package SRT uses a function **atsmeans(.)** to estimate AI values. There are several options provided by this function that allow users to customize the outputs based on

the interests of studies. By specifying the option of method="MLE" or method="IPW", I can choose to use maximum likelihood estimator or inversed probability weighting estimator for estimation. The default of this option is MLE. This function also provides an logic option "common" so that users can choose to apply the common variance across all the treatment sequences for estimation. This function also gives an logic option "conf" for users to choose whether the output of estimated AI values are with confidence intervals. Users can control the confidence levels of these intervals at $(1-\alpha)\%$ level by specifying the numeric option "alpha". The function atsmeans(.) outputs 3 parts of contents. First, it outputs the value matrix, in which each row represents one AI. Therefore, for a SMART with total number of G AIs embedded in, the value matrix always have G rows. For each row, the decision makings under the corresponding AI are listed and it also shows total number of patients following the AI, estimated AI value, standard error of estimation and confidence intervals (if specify "CI=True"). This function also automatically outputs the estimated variance-covariance matrix for the estimated values of all the AIs embedded in SMART. For a SMART with G AIs, the variance-covariacne matrix has a dimension of $G \times G$. The functions provides a logic option "plot". If users specify "plot=True", it output a figure of all the estimated AI values with $(1-\alpha)\%$ level confidence intervals.

Figure 4.5. Estimated AI values by atsmeans(.)

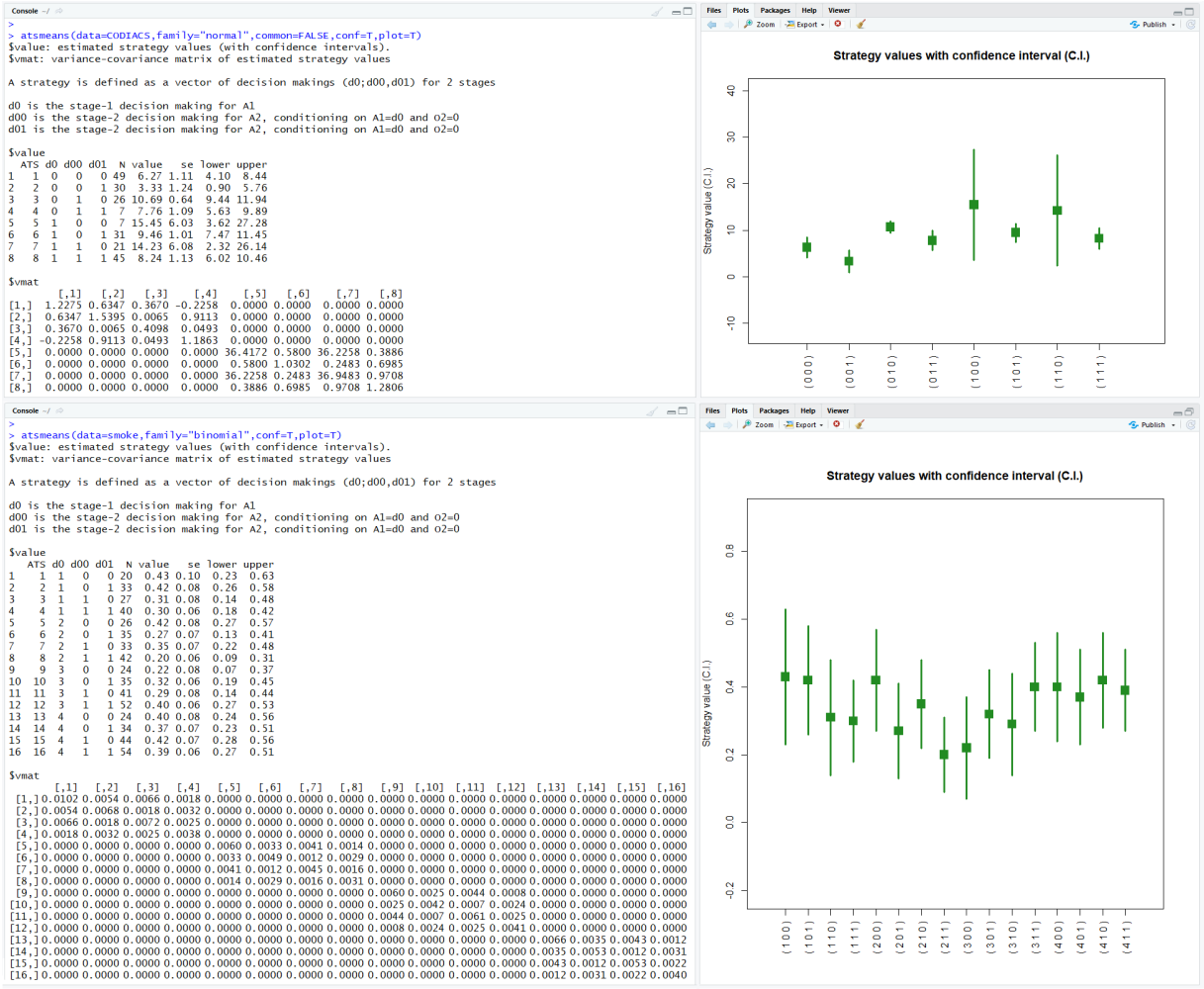


Figure 4.5 shows the outputs of the function atsmeans(.) using the two example data mentioned before. The upper-left window is the program and output of calling the function using the depression management trial data. There were up to 8 AIs embedded in this SMART data. I use $g = 1, \ldots, 8$ as the index of AI. For example, there were 49 patients in this dataset followed the AI with $g = 1$. The estimated value of this AI is 6.27 with 95% confidence interval $(4.10, 8.44)$. The plot of these AI values is shown on the upper-right window. The estimated variance-covariance matrix is shown below the value matrix. We can see that the estimated variances corresponding to AIs with $g = 5$ and $g = 7$ are larger than others. The covariances between AIs with different stage-

1 treatment are consistently equal to zero. The lower-left window is the program and output of using the smoking cessation data, which has binary primary outcome. When we specify the option "family=binomial", the calculation will automatically assuming uncommon variance across all the treatment sequences. The 16×16 estimated variance-covariance matrix is also given in the output. We can see that the block diagonal matrices are non-zero, while others are all zeros. The value plots of this example is shown in the lower-right window, with the observed best AI with the index $g = 1$ being the one has greatest estimated value and $g = 8$ with the smallest estimated value.

4.5 Comparing Multiple AIs

SMART is an efficient design for early phase clinical trials. The goals of these studies are usually providing information about selecting one promising or several near-best AIs and moving forward to further clinical investigation, in which a confirmatory trial will be conducted by comparing the selected AI(s) with a appropriate reference intervention. In these situations, the key of statistical analysis is comparing the values of multiple AIs embedded in SMART. Several analytical methods can be applied to do this job, such as the omnibus test mentioned in Chapter 2.2 and pairwise comparison testing procedure described in Chapter 2.3. Another technique can be building a series of simultaneous confidence intervals, such as the MCB confidence intervals introduced in Chapter 3. The package SRT uses a function called **smartest(.)** to make such comparisons.

Same as the functions of seqmeans(.) and atsmeans(.), the function smartest(.) also provides options "family" for users to specify the type of primary outcome, "method" to select the estimation method, and "common" to choose whether the common variance

assumption across all treatment sequences should be applied. This function automatically outputs three parts of results, which are the list of all AIs, results of the global test, and pairwise comparisons. First, it gives a list of AI under "$Strategy", which provides the information of all the AI embedded in the input SMART data and the decision makings under each AI. The total number of patients under each AI is also given in the last column of this section. In the second section of output, this function gives the results of a global test with the hypothesis

$$H_0 : \mu_1 = \cdots = \mu_G \text{ versus } H_1\text{: } \mu_g\text{'s are not all equal for } g = 1, \ldots, G.$$

For a SMART consists of a large number of AIs, such a global test can be applied as the gate-keeping test to avoid exhaustive search, so as to control the familywise type I error. The details of conducting such a global test is described in Chapter 2.2 and Chapter 2.3. Under the output "$Global.test", it gives the information about the total number of patients participated in a SMART, total number of AIs embedded in the SMART data, degrees of freedom, test statistics and P-value of this test. The third part of output gives a matrix of the results of pairwise comparisons. For a SMART data include G AIs, there are $G \times (G-1)$ pairwise comparisons. Each row of this section corresponds to one pairwise comparison and it provides information of the indexes of two AIs, the difference of two AI values with a $(1-\alpha)\%$ level confidence interval, the test statistics and the P-value of the pairwise comparison with the hypothesis

$$H_0 : \mu_j = \mu_k \text{ versus } H_1\text{: } \mu_j \neq \mu_k, \text{ where } j, k = 1, \ldots, G \text{ and } j \neq k.$$

The function smartest(.) provides an option "adjust", by which users can control the type

of confidence intervals given in the output. The default of this option is "adjust=n", which gives the confidence intervals without adjusting for multiplicity. If specify "adjust=Bon", the function will return confidence intervals adjusted for Bonferroni correction. By default, for a SMART data including G AIs, the confidence intervals will adjusted for $\frac{G\times(G-1)}{2}$ pairs of comparisons. In practice, users can control the number of pairwise comparisons in multiplicity adjustment by specifying a value to an option named "npairs=". Users can also choose "adjust=MCB", which leads to MCB intervals in the output. Since MCB confidence intervals are comparing each AI value with the truly best AI, which is assumed to be unknown in analysis, for a SMART data with G AIs, there are G MCB confidence intervals. Consequently, When the "adjust=MCB" is specified, there are only G rows in this section of output and each row corresponds to one AI.

Figure 4.6 shows the outputs of comparing AIs embedded in the smoke cessation trial data based on the function smartest(.). There are 4 options for stage-1 treatment. From the first part of output, users can see that there are 16 AIs and the decision makings under each AI. Each stage-1 treatment option corresponds to 4 AIs. The second part gives the results of the global test. It shows that there are 282 patients who completed this SMART and totally 16 AIs embedded in the study design. The test statistics of this global test calculated based on the sample data is 19.14 and it follows a chi-squared distribution with 11 degrees of freedom under the null hypothesis of no difference among all the AI values, which leads to a P-value equal to 0.0587. Therefore, I conclude that the AI values are not statistically different among all the 16 AIs embedded in the study at 5% nominal significance. The third part of the output gives the results of pairwise comparisons. For example, the first row gives the difference of values of AIs $g = 1$ vs. $g = 2$, where $g = 1$ is the AI of $(1; 0, 0)$ and $g = 2$ is the AI of $(1; 0, 1)$. This part of output shows that the value

difference is 0.01 with a 80% confidence intervals $(-0.24, 0.25)$ adjusted for Bonferroni correction. Since no value of "npair" is specified, the simultaneous intervals is adjusted for $\frac{16\times15}{2} = 120$ pairwise comparisons. There are $16 \times 15 = 240$ rows in this section of output. Figure 4.7 is the output of "$pairwise.comparisons" when choose "adjust=MCB" using the depression management trial data. In this case, I am not comparing two observed AI values, but comparing the value of each AI with the unknown truly best AI. There are totally 8 rows in these output and the MCB interval of AI with $g = 2$ has an upper limit with negative value. Therefore, we identify this AI as inferior to the truly best AI.

Figure 4.6. Compare multiple AIs using the smoking cessation trial data by smartest(.)

```
Console ~/
>
> smartest(data=smoke,family="binomial",alpha=0.2,adjust="Bon")
$Strategy provides the details of decision makings under strategy labels (ATS)
$Global.test assesses the null hypothesis of no difference across all the strategy values
$Pairwise.test compares all the pairs of strategies, of which the labels are shown in $Strategy.

$Strategy
      ATS d0 d00 d10  N
 [1,]   1  1   0   0 20
 [2,]   2  1   0   1 33
 [3,]   3  1   1   0 27
 [4,]   4  1   1   1 40
 [5,]   5  2   0   0 26
 [6,]   6  2   0   1 35
 [7,]   7  2   1   0 33
 [8,]   8  2   1   1 42
 [9,]   9  3   0   0 24
[10,]  10  3   0   1 35
[11,]  11  3   1   0 41
[12,]  12  3   1   1 52
[13,]  13  4   0   0 24
[14,]  14  4   0   1 34
[15,]  15  4   1   0 44
[16,]  16  4   1   1 54

$Global.test
  size nATS df chisq Pvalue
1  282   16 11 19.14 0.0587

$Pairwise.test
       label  diff lower.CI upper.CI     Z Pvalue
  1   1 vs. 2  0.01    -0.24     0.25  0.11 0.9133
  2   1 vs. 3  0.12    -0.08     0.33  1.90 0.0574
  3   1 vs. 4  0.13    -0.19     0.45  1.30 0.1938
  4   1 vs. 5  0.01    -0.39     0.41  0.05 0.9616
  5   1 vs. 6  0.16    -0.23     0.55  1.31 0.1914
  6   1 vs. 7  0.08    -0.30     0.46  0.64 0.5249
  7   1 vs. 8  0.23    -0.13     0.59  2.01 0.0443
  8   1 vs. 9  0.21    -0.19     0.61  1.67 0.0959
  9   1 vs.10  0.11    -0.27     0.49  0.91 0.3626
 10   1 vs.11  0.14    -0.27     0.54  1.06 0.2898
 11   1 vs.12  0.03    -0.34     0.41  0.27 0.7868
 12   1 vs.13  0.03    -0.38     0.43  0.20 0.8406
 13   1 vs.14  0.06    -0.33     0.45  0.48 0.6298
 14   1 vs.15  0.01    -0.38     0.40  0.04 0.9648
 15   1 vs.16  0.04    -0.33     0.41  0.33 0.7407
...      ...   ...       ...      ...   ...    ...

226  16 vs. 1 -0.04    -0.41     0.33 -0.33 0.7407
227  16 vs. 2 -0.03    -0.36     0.30 -0.30 0.7662
228  16 vs. 3  0.08    -0.25     0.42  0.79 0.4292
229  16 vs. 4  0.09    -0.19     0.37  1.04 0.2964
230  16 vs. 5 -0.03    -0.35     0.28 -0.33 0.7399
231  16 vs. 6  0.12    -0.18     0.42  1.28 0.2008
232  16 vs. 7  0.04    -0.25     0.33  0.41 0.6842
233  16 vs. 8  0.19    -0.07     0.46  2.28 0.0227
234  16 vs. 9  0.17    -0.14     0.49  1.72 0.0855
235  16 vs.10  0.07    -0.22     0.36  0.77 0.4424
236  16 vs.11  0.10    -0.22     0.41  0.95 0.3424
237  16 vs.12 -0.01    -0.29     0.28 -0.08 0.9368
238  16 vs.13 -0.01    -0.30     0.27 -0.15 0.8816
239  16 vs.14  0.02    -0.16     0.20  0.36 0.7181
240  16 vs.15 -0.03    -0.25     0.19 -0.49 0.6271
>
```

Figure 4.7. MCB intervals using the CODIACS data by function smartest(.)

```
Console ~/
> smartest(data=CODIACS,family="normal",adjust="MCB",alpha=0.2)
$MCB confidence interval
  ATS value  lower upper
1   1  6.27 -19.69  0.00
2   2  3.33 -22.68 -0.26
3   3 10.69 -15.15  0.00
4   4  7.76 -18.20  0.00
5   5 15.45  -7.61  0.00
6   6  9.46 -16.31  0.00
7   7 14.23  -8.92  0.00
8   8  8.24 -17.62  0.00
>
```

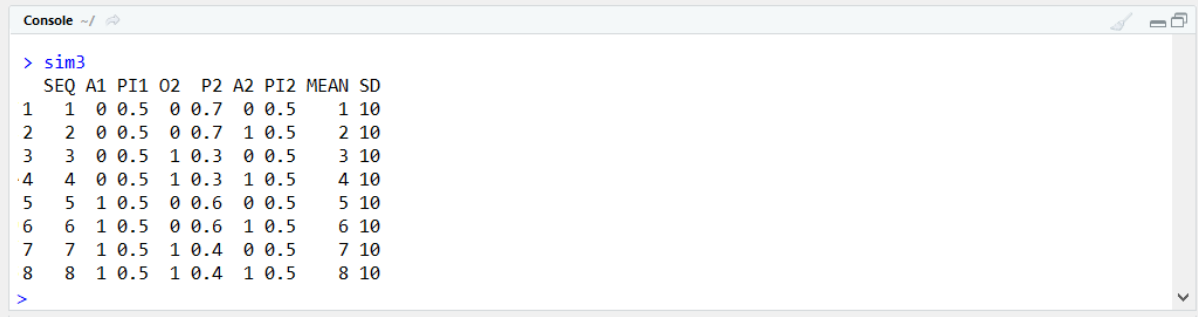
4.6 Sample Size Calculation

The goal of sample size calculation in planning a SMART is to estimate the minimal required sample size that can help to achieve the targeted comparative power under the rigorous control of the pre-specified error rate. In practice, the sample size calculation is based on the planned analytical method decided at the design stage. The package SRT provides a function, **smartsize(.)**, to conduct the sample size calculation.

This function allows to choose whether to conduct the sample size calculation based on a global test, or a pairwise test between two pre-specified AIs. Users can use the logic option "global=" to control the test by which the sample size calculation is conducted. The default is "global=True". By specifying the values of two options "alpha=" and "beta=", users can control the targeted type I error rate at α and targeted power at $(1-\beta)\times 100\%$. The default of these two options are "alpha=0.05" and "beta=0.20". When users choose "global=True", there are two choices to input the required information for sample size calculation. First, users can directly input the standardized effect size (Δ) via the option "delta=" and the degrees of freedom of chi-squared test (ν) via the option "df=". The details of how to calculate the effect size and degrees of freedom

are described in (9) and (7) in Chapter 2.2. Alternatively, investigators can input the sequence information matrix (SIM) by an option "sim=", which is a data frame contains all the sequence-specific information as shown in Figure 4.8. Each row of SIM represents a treatment sequence in SMART. The columns of SIM include the index of sequence (SEQ), the baseline and intermediate outcomes variables ($O1, O2, \ldots, OT$), the response rates ($P1, P2, \ldots, PT$), the stage-specific treatments($A1, A2, \ldots, AT$), the randomization probabilities ($PI1, PI2, \ldots, PT$), and the targeted sequence-specific mean (MEAN) and standard deviation (SD).

Figure 4.8. Sequence information matrix (SIM)



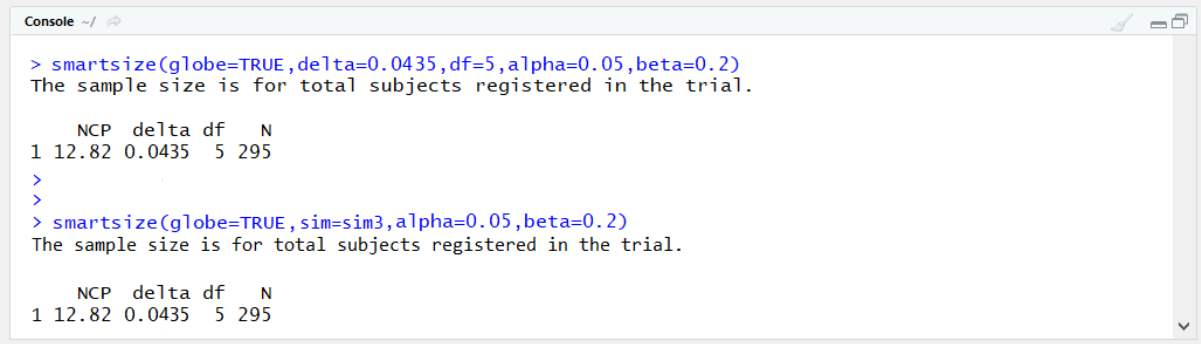
```
Console ~/

> sim3
  SEQ A1 PI1 O2  P2 A2 PI2 MEAN SD
1   1  0 0.5  0 0.7  0 0.5    1 10
2   2  0 0.5  0 0.7  1 0.5    2 10
3   3  0 0.5  1 0.3  0 0.5    3 10
4   4  0 0.5  1 0.3  1 0.5    4 10
5   5  1 0.5  0 0.6  0 0.5    5 10
6   6  1 0.5  0 0.6  1 0.5    6 10
7   7  1 0.5  1 0.4  0 0.5    7 10
8   8  1 0.5  1 0.4  1 0.5    8 10
>
```

Figure 4.9 is the R inputs and outputs of using smartsize(.) to calculation the sample size for a SMART. The sequence information matrix (sim3) is shown in Figure 4.8. This is a 2-stage SMART with stage-1 treatment not adapted to baseline information. Thus, the stage-specific treatment and intermediate response can be summarized by (A_1, O_2, A_2). There are 8 treatment sequences embedded in the design. The targeted mean and the targeted standard deviation for each sequence are specified in the last two columns of the matrix. The effect size calculated based on sim3 is $\Delta = 0.0435$ and the degrees of freedom of chi-squared test is df=5. By inputing the targeted effect size and degrees of freedom into smartsize(.), I knew from the R output that the total sample size of 295 patients will help to achieve 80% power. I also directly input the sequence information matrix (sim3),

and it gave the same results. In practice, sequence information matrix is helpful when statisticians communicate with clinical investigators. It displays the clinical history of each type of patients during the trial in a straightforward manner, so investigators can have better sense in setting up the targeted means and standard deviations at design stage.

Figure 4.9. Sample size calculation by smartsize(.)



In some case, the investigator is particularly interest in comparing a pair of AIs. By specifying the option "global=FALSE", the function smartsize(.) will return the sample size calculated based on a pairwise test and return the total numbers of patients included in the pair of AIs of interest. The total sample size of SMART can be obtained by adjusting the size of pairwise test based on the randomized probabilities determined at design stage. In that situation, the sample size calculation is similar to a traditional t test.

Chapter 5 An Exploratory Study to Design a SMART for Comparing Multiple Patient Care Strategies for Depression Management

5.1 Introduction

I explored the distribution theory of maximum likelihood estimators (MLEs) of adaptive intervention (AI) values under general SMART designs and proposed a Wald test for overall equality based on MLEs in Chapter 2. Such a test can be applied as a gatekeeping test for comparing multiple AIs embedded in a SMART. Another important contribution of the proposed test is that we can use the formal sample size calculation formula and power calculation formula derived based on the proposed Wald test to design a SMART. In this chapter, I will show an example about how to design a SMART for comparing multiple patient care strategies against depression.

It is common for patients to suffer from depression after the surgeries for acute coronary syndrome (ACS) (Bush et al., 2005). A comprehensive review noted that there are almost 2 out of every 5 post-ACS patients have clinically significant depression (Carney and Freedland, 2008), which has been reported to be observationally associated with diminished health-related quality of life (Stafford et al., 2007), high costs (Von Korff et al., 1992) and poor medical prognosis (Nicholson, Kuper and Hemingway, 2006). Consequently, a patient care strategy, named *screen and treat*, which recommends administering a depression screening questionnaire to post-ACS patients and referring those who are positive in the screen for depression treatment, has been incorporated into the guidelines from the American Heart Association (AHA) and endorsed by multiple societies (Lichtman et al., 2008; Graham et al., 2007; National Institute for H, Clinical E, 2009).

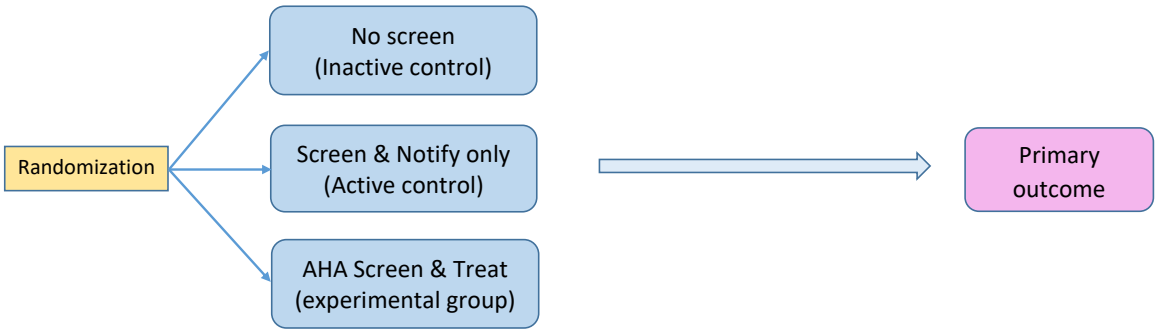
However, there is no randomized clinical trial (RCT) to inform such a large scale and potentially expensive patient care strategy. Evidence-based clinical guidelines typically review all available research on a specific topic systematically and make recommendations by integrating such information. Comparing to observational studies, RCTs are usually more replicable, have fewer sources of bias and strong ability in controlling statistical power and error, and thus are given greater emphases as evidences for practice guidelines (Guyatt et al., 2000). The study in this chapter was motivated by the needs for designing a RCT to evaluate the net benefit of AHA screen and treat strategy and compare it with other two common post-ACS patient care strategies. As the AHA screen and treat strategy involves two clinical actions given sequentially over a period of time and the subsequent action (i.e., referral to treat or not) taken by a patient depends on the result of the first action (i.e., depression screen) within the same individual, a patient under this strategy will follow one of two clinical sequences, "depression screen $\rightarrow$ negative $\rightarrow$ notification" and "depression screen $\rightarrow$ positive $\rightarrow$ referral to treat". Thus, it fits the adaptive intervention paradigm. See Chapter 1 for more details about AI definition and Cheung et al. (2015) for an example of AI in depression management. Specifically, I am interested in designing a RCT for comparing the AHA depression screen and treat strategy (experimental strategy) with the no screening strategy (inactive control), because no RCT on depression screen for post-ACS patients has been conducted. In addition, we are also interested in a minimal enhanced strategy, called *screen and notify* (active control), which provides all post-ACS patients depression screen and notifies the appropriate in-network primary care provider (PCP) of the existence of depression. The primary outcome considered for this study is the improvement of quality-adjusted life year (QALY) at 18 months after baseline. A strategy leading to a greater improvement of QALY is

viewed as a better patient care strategy.

5.2 Study Designs: SMART vs. RAB

I illustrate in this section two different clinical trial designs with the goal to compare the 3 patient care strategies mentioned in Chapter 5.1. A clinical trial named Comparison of Depression Interventions after Acute Conorary Syndrome - Quality of Life (CODIACS-QOL), was originally proposed to compare the three patient care strategies of interest. Based on the design of this trial proposed in protocol, patients were randomly assigned to 3 arms at baseline with equal probabilities. Each treatment arm corresponded to one patient care strategy as shown in Figure 5.1. Such a randomized at baseline (RAB) design helps to rule out the possible impacts from unmeasurable confounders. As a result, data collected using RAB design allow to make comparisons among 3 strategies with minimal bias; and in principle, the results could be referred for evidence-based practice guideline. Those patients who are screen negative in both the experimental group and the active control group would have the exactly same clinical experience (i.e., depression screen $\rightarrow$ negative $\rightarrow$ notification) during the study, and thus it is reasonable to expect that the primary outcomes of these patients are identically distributed. However, under such a design, only a portion of patients who followed this sequence contributed in estimating the effect of either the experimental strategy or the active control strategy, which can possible lose efficiency.

Figure 5.1. Design diagrams for comparing 3 patient care strategies after ACS: Randomized at baseline (RAB)



Alternatively, we can design a SMART to compare the 3 strategies of interest as shown in Figure 5.2. Under the SMART design, a patient is initially randomized to receive depression screen or no screen, and then re-randomized at the subsequent stage to receive treatment directly or to be notified about the existence of depression. By virtue of sequential randomization, the assumption of ignorable treatment holds and thus the results can be referred as RCT-based evidence for practice guideline. Patients who complete the SMART design can possibly follow 4 clinical sequences: (1) no depression screen, (2) depression screen $\rightarrow$ negative $\rightarrow$ notification (3) depression screen $\rightarrow$ positive $\rightarrow$ notification and (4) depression screen $\rightarrow$ positive $\rightarrow$ referral to treat. Comparing to the RAB design, data collected from all the patients who screen negative in SMART contribute to evaluate both the experimental strategy and the active control strategy, and thus can potentially improve the design efficiency.

It is important to note that this SMART design involves AIs with different types of structures, featured by 3 parameters: the number of stages, the number of action options given clincial history, and the number of intermediate response categories. This is different from the traditional SMART research focusing on comparing AIs with the

same structure (Robins, 1986; Lavori et al. 2007; Murphy et al. 2001). I have shown in Chapter 2.3 that the proposed Wald test is valid for SMART designs with varying design structures and randomization schemes, and thus it can be applied to the design shown in Figure 5.2. However, whether we can design a SMART trial to improve some desired features, such as power or average patient care, comparing to the RAB design proposed in original protocol, remains unclear. In this chapter, I explore various combinations of randomization probabilities that yield different powers of SMART designs.

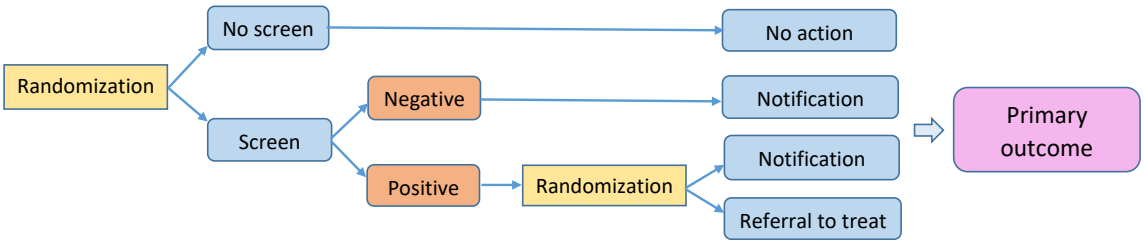


Figure 5.2. Design diagrams for comparing 3 patient care strategies after ACS: Sequential Multiple Assignment Randomized Trial (SMART)

5.3 Notation and Method

5.3.1 Notation

I use the same notations under the general SMART design as described in Chapter 2.2.1. Given the design structure shown in Figure 5.2, let T_1 and T_2 denote without depression screen and with screen, respectively, at Stage 1. Under T_2, there are two possible response categories, R_{21} for screen negative and R_{22} for screen positive. While under T_1, all the patients have the same intermediate outcome R_{11}. Given the clinical history of (T_2, R_{22}),

patients are randomized to two Stage-2 actions, S_{221} for notification and S_{222} for referral to treat. Also, I use S_{111} and S_{121} to represent the Stage-2 action for those with history (T_1, R_{11}) and (T_2, R_{21}). Specifically in the data, I code 1 for depression screen and 0 for no screen. Also, I code 1 for referral to treat and 0 for notification for those who screen positive at Stage 2. I use 0 to represent screen negative and 1 screen positive. In addition, those who receive no screen do not have any screen result (coded as 0) or action (coded as 0) at Stage 2. Thus, the design structure of SMART can be summarized as

$$(T_1,\ T_2) = (0,\ 1)$$

$$(R_{11},\ R_{21},\ R_{22}) = (0,\ 0,\ 1)$$

$$(S_{111},\ S_{211},\ S_{221},\ S_{222}) = (0,\ 0,\ 0,\ 1).$$

Let (π_1, π_2) be the Stage-1 randomization probabilities corresponding to (T_1, T_2). Let (π_{221}, π_{222}) be the Stage-2 randomization probabilities for (S_{221}, S_{222}). Considering that both (π_1, π_2) and (π_{221}, π_{222}) sum up to 1, given the design structure in figure 4.2, the SMART design is completely depicted by (π_2, π_{222}). I denote the screen and notify strategy by $(1; 0, 0)$ and the AHA screen and treat strategy by $(1; 0, 1)$. Also, under the SMART framework, we can view the no screen strategy as a degenerate case of AI, and thus denote it by $(0; 0, 0)$.

5.3.2 Analytical Methods

The CODIACS-QOL protocol proposed to compare three strategies by pairwise testing procedure. Specifically, t-test with Bonferroni adjustment was used to compare each pair of strategies. Since this is a classic testing procedure used in traditional RCT, I skip the

details of the test. For SMART design, I apply the likelihood-based Wald test described in Chapter 2.2. Let $\boldsymbol{\Theta} = (\theta_1, \theta_2, \theta_3)$ be the values of 3 AIs (i.e., no screen, screen and treatment, screen and treat) embedded in the study. A Wald test is proposed with the following hypotheses:

$$H_0 : \theta_1 = \theta_2 = \theta_3 \text{ versus } H_1\text{: } \theta_g\text{'s are not all equal for } g = 1, 2, 3. \tag{18}$$

Let n be the total sample size of a SMART. Let

$$\boldsymbol{C} = \begin{pmatrix} 1 & -1 & 0 \\ 1 & 0 & -1 \end{pmatrix}$$

be the contrast matrix. Let $\hat{\boldsymbol{\Theta}}$ be the MLEs of $\boldsymbol{\Theta}$ and $\hat{\boldsymbol{\Sigma}}$ be the estimated asymptotic covariance matrix of $\hat{\boldsymbol{\Sigma}}$. A Wald-type test statistic can be written as

$$Q = n(\boldsymbol{C}\hat{\boldsymbol{\Theta}})^T(\boldsymbol{C}\hat{\boldsymbol{\Sigma}}\boldsymbol{C}^T)^-(\boldsymbol{C}\hat{\boldsymbol{\Theta}}), \tag{19}$$

where $\boldsymbol{M}^-$ denotes a generalized inverse of a square matrix $\boldsymbol{M}$. Under the null hypothesis of (18), the test statistics $Q \xrightarrow{d} \chi^2_\nu$, a chi-squared distribution with degrees of freedom

$$\nu = \sum_{i=1}^{I}\sum_{j=1}^{J_i} K_{ij} - \sum_{i=1}^{I} J_i + I - 1 = 4 - 3 + 2 - 1 = 2. \tag{20}$$

While, under the alternative hypothesis of (18), $Q \xrightarrow{d} \chi^2_\nu(\lambda^*)$, a noncentral chi-squared distribution with noncentrality parameter λ^*. I have shown in Chapter 2.3 that this test has nice asymptotic properties and performances under SMART designs with varying design structures and randomization schemes. More details of the test see Chapter 2.2.

5.3.3 Power Calculation

The power of the Wald test given the targeted response probabilities $\{p_{ij}\}$ and targeted sequence-specific outcome parameters $\{\phi_{ijk}, \tau_{ijk}\}$ can be calculated as follows:

Step 1: Given the design structure $\{T_i, R_{ij}, S_{ijk}\}$ and a pre-specified type I error rate α, get the critical value (CV) of rejecting H_0 in (18) by

$$CV = \chi^2_{\nu,\alpha}.$$

Step 2: Given total sample size n, randomization probabilities $\{\pi_i, \pi_{ijk}\}$ and targeted $\{p_{ij}, \phi_{ijk}, \tau_{ijk}\}$, calculate the non-centrality parameter λ^* by

$$\lambda^* = n\big(\boldsymbol{C}\boldsymbol{\Theta}^*\big)^T\big(\boldsymbol{C}\boldsymbol{\Sigma}^*\boldsymbol{C}^T\big)^-\big(\boldsymbol{C}\boldsymbol{\Theta}^*\big),$$

where $\boldsymbol{\Theta}^*$ is the targeted AI values and $\boldsymbol{\Sigma}^*$ is the targeted asymptotic covariance matrix of MLEs.

Step 3: Given ν, λ^* and CV, calculate the power by

$$\text{Power} = 1 - \Pr(\chi^2_{\nu,\lambda^*} \leq CV). \tag{21}$$

Note that $\boldsymbol{\Sigma}^*$ is a function of $\{\pi_i, \pi_{ijk}\}$ given $\{T_i, R_{ij}, S_{ijk},\}$ and $\{p_{ij}, \phi_{ijk}, \tau_{ijk}\}$, so that λ^* is a function of $\{\pi_i, \pi_{ijk}\}$, and thus, the power of Wald test given $\{T_i, R_{ij}, S_{ijk},\}$ and $\{p_{ij}, \phi_{ijk}, \tau_{ijk}\}$ is a function of $\{\pi_i, \pi_{ijk}\}$.

5.4 Compare Designs by Numerical Computation

In this section, I compare SMART versus RAB by numerical computation. The goal is to explore the possibility of improving some desired features of RCT by replacing the original proposed RAB design with a SMART design. Specifically, I compare several SMART designs with different randomization probabilities (π_2, π_{222}) versus the RAB design proposed in CODIACS-QOL protocol under difference outcome scenarios in terms of type I error rate, comparative power, and average primary outcome of all the trial participants.

5.4.1 Outcome Scenarios

I considered 9 outcome scenarios using 3 screen positive rates $P = (0.1, 0.2, 0.3)$ and 3 value patterns (VP) of targeted treatment sequence-specific means suggested by the CODIACS-QOL protocol. The final primary outcome Y_{ijk} given the clinical history of (T_i, R_{ij}, S_{ijk}) was randomly generated based on a normal distribution with sequence-specific means

$$\text{VP1: } (\phi_{111}, \phi_{211}, \phi_{221}, \phi_{222}) = (0.055, 0.055, 0.055, 0.21)$$

$$\text{VP2: } (\phi_{111}, \phi_{211}, \phi_{221}, \phi_{222}) = (0.055, 0.055, 0.080, 0.21)$$

$$\text{VP3: } (\phi_{111}, \phi_{211}, \phi_{221}, \phi_{222}) = (0.055, 0.055, 0.130, 0.21)$$

and sequence-specific variance $\sigma^2 = 0.17^2$. ϕ_{ijk} was specified by

$$\phi_{ijk} = \beta_0 + \beta_1 T_i + \beta_2 R_{ij} + \beta_3 S_{ijk} + \beta_4 T_i R_{ij} + \beta_5 T_i S_{ijk} + \beta_6 R_{ij} S_{ijk} + \beta_7 T_i R_{ij} S_{ijk}$$

for $T_i, R_{ij}, S_{ijk} \in \{0, 1\}$, and the details of choice β's see Appendix A2. Under VP1, the depression screen itself has no impact on primary outcome so that the post-ACS patients under the no screen strategy have the same expected outcome as those under the screen and notify strategy. In other words, the improvement of primary outcome for those patients under the AHA screen and treat strategy is completely due to the action of referral to treat. Under VP2, the screen and notify strategy has mild positive impact on the primary outcome, comparing to those under no screen strategy. But the improvement of outcome is mainly driven by the action of referral to treat. Under VP3, notifying patients the existences of depression has strong impact on primary outcome. Figure 5.3 shows the AI values in all 9 outcome scenarios, with each row corresponding to a value pattern and each column corresponding to a screen positive rate.

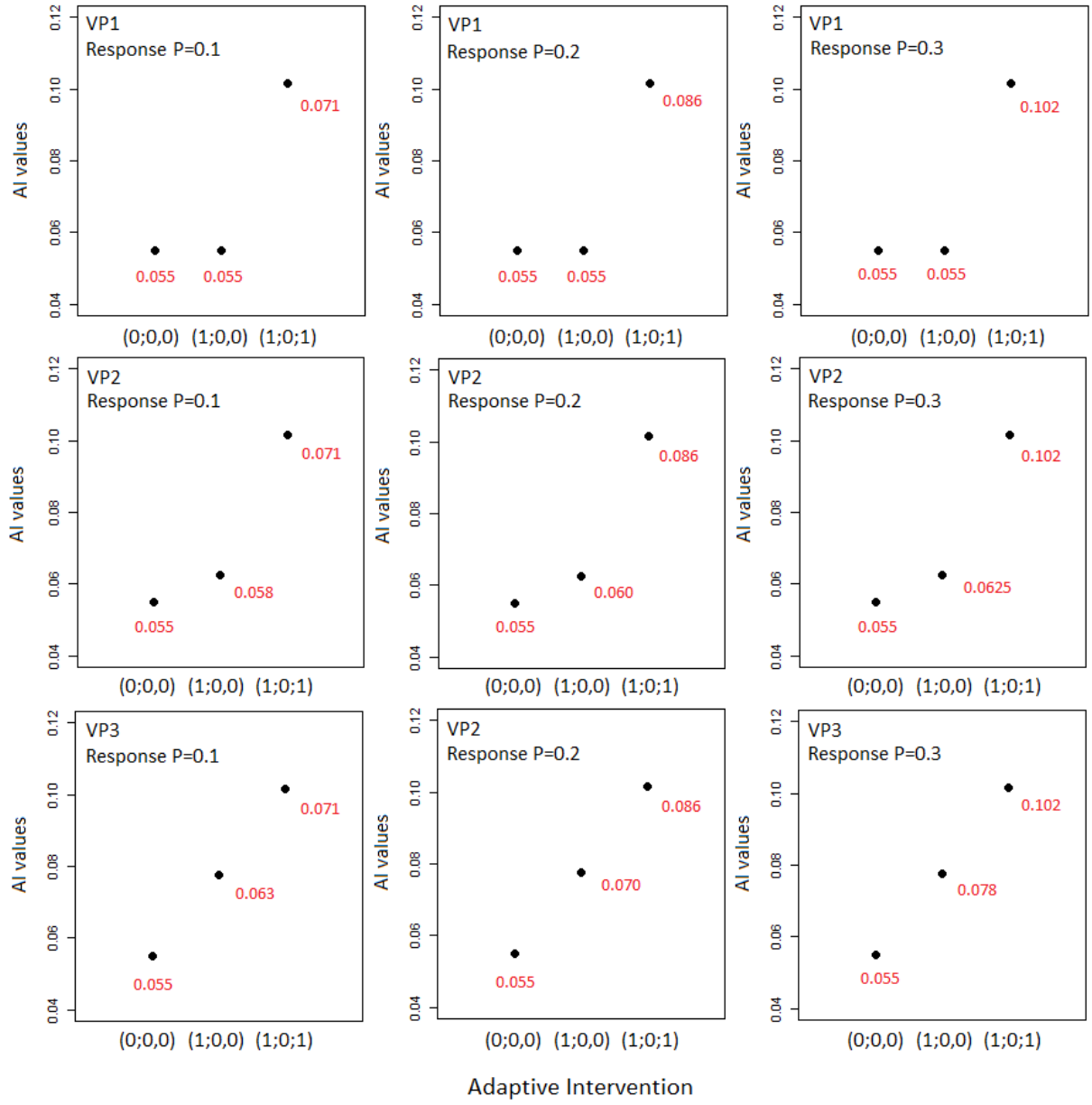


Figure 5.3. Outcome scenario of AIs considered in the simulation. (0;0,0) is no screen strartegy; (1;0,0) is screen and notify strategy; (1;0,1) is screen and treat strategy.

5.4.2 SMART Design

For each outcome scenario, I calculated the theoretical powers by (21) under each pair of (π_2, π_{222}) with values varying from 0.01 to 0.99. The type I error rate was set to be $\alpha = 0.05$ and the total sample size was fixed at $n = 400$. Figure 5.4 gives the surface

plots of theoretical powers under each pair of (π_2, π_{222}). Under VP1 and VP2, given a certain value of π_{222}, the power increases as π_2 increases. While given π_2, $\pi_{222} = 0.5$ achieves the maximum power. Under VP3, when the screen positive rate is $P = 0.1$, the power varies by (π_2, π_{222}) in the same style as under VP1 and VP2. When $P = 0.2$, the randomization probabilities π_2 with a value between 0.6 to 0.8 results in a greater power than other values. When $P = 0.3$, π_2 with a value between 0.5 and 0.7 results in a greater power than other values. I selected 5 SMART designs (S1-S5) under each outcome scenario as shown in Table 5.1. For VP1 and VP2, π_{222} was fixed at 0.5 and π_2 took a value of 0.5-0.9 with 0.1 increment in S1-S5. For VP3, with $P = 0.1$, I used the same designs as VP1 and VP2; while with $P = 0.2$ or 0.3, I chose 5 combinations of π_2 and π_{222} values that can yield high power as depicted in Figure 5.3. I evaluated the actual type I error rates, empirical powers and average primary outcomes of these SMART designs and compared them with the RAB design described in Chapter 5.2.1 using simulation. For each simulated SMART data, I conducted a proposed Wald test as described in Chapter 2.2. For each simulated RAB data, I conducted pairwise t tests with Bonferroni adjustment. All the simulation results were generated based on 5000 replicates.

Value Pattern	Screen Positive Rate	S1	S2	S3	S4	S5
VP1	P=0.1	(0.5,0.5)	(0.6,0.5)	(0.7,0.5)	(0.8,0.5)	(0.9,0.5)
	P=0.2	(0.5,0.5)	(0.6,0.5)	(0.7,0.5)	(0.8,0.5)	(0.9,0.5)
	P=0.3	(0.5,0.5)	(0.6,0.5)	(0.7,0.5)	(0.8,0.5)	(0.9,0.5)
VP2	P=0.1	(0.5,0.5)	(0.6,0.5)	(0.7,0.5)	(0.8,0.5)	(0.9,0.5)
	P=0.2	(0.5,0.5)	(0.6,0.5)	(0.7,0.5)	(0.8,0.5)	(0.9,0.5)
	P=0.3	(0.5,0.5)	(0.6,0.5)	(0.7,0.5)	(0.8,0.5)	(0.9,0.5)
VP3	P=0.1	(0.5,0.5)	(0.6,0.5)	(0.7,0.5)	(0.8,0.5)	(0.9,0.5)
	P=0.2	(0.8,0.5)	(0.8,0.6)	(0.7,0.7)	(0.65,0.8)	(0.6,0.9)
	P=0.3	(0.7,0.5)	(0.7,0.6)	(0.65,0.7)	(0.6,0.8)	(0.55,0.9)

Table 5.1. Randomization probabilities (π_2, π_{222}) considered in each outcome scenario

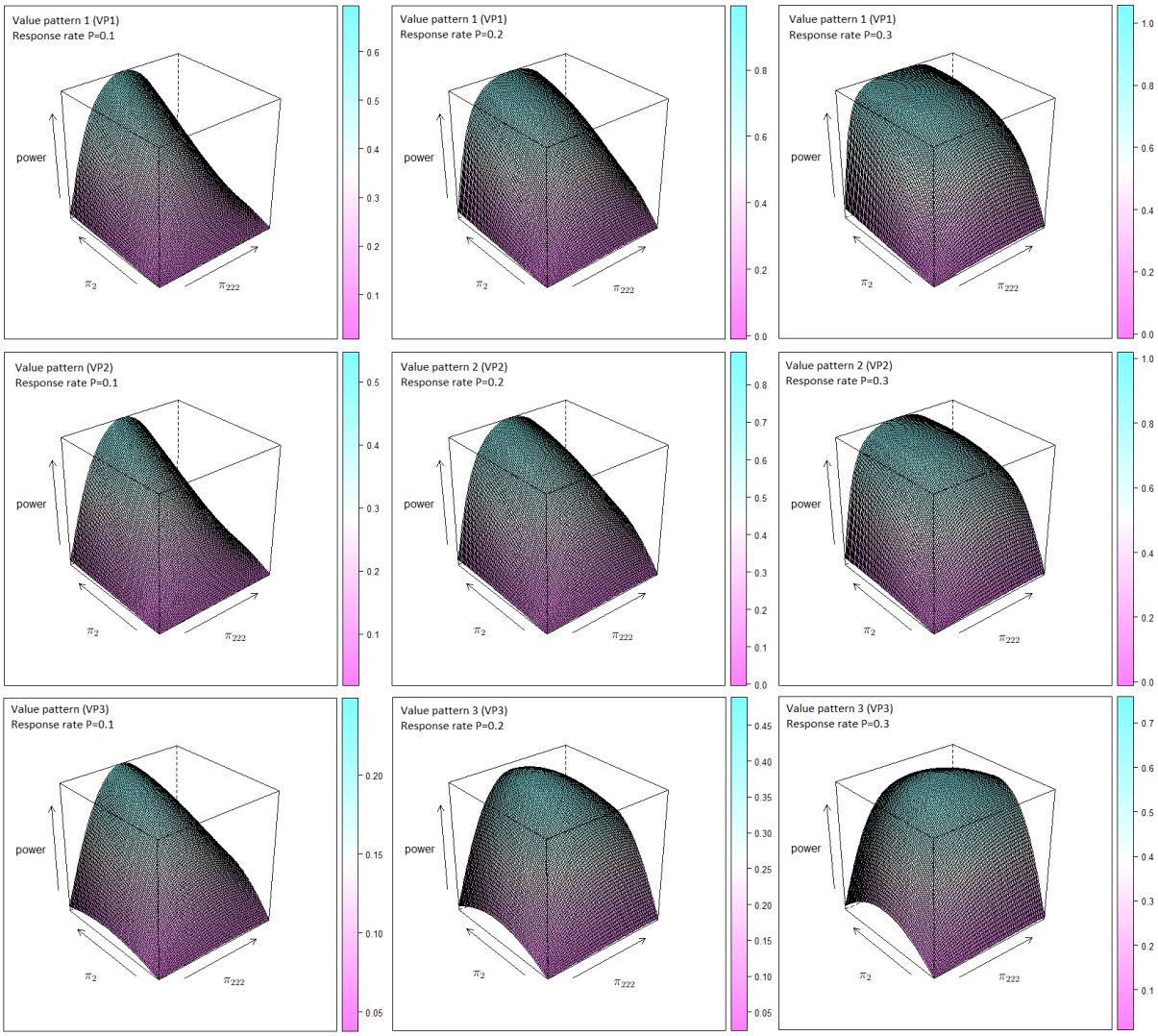


Figure 5.4. Theoretical powers of the Wald test by randomization probabilities.

5.4.3 Comparison Results

Table 5.2 gives the actual type I error rates of applying the Wald test to SMART designs (S1-S5) at 5% nominal level. Each row in this table corresponds to an outcome scenario described in Chapter 5.4.1. All these actual type I error rates are fairly close to the nominal level of 5%. For comparison purpose, I also considered pairwise testing procedures comparing 3 strategies under the RAB design, with or without multiplicity adjustment.

The pairwise testing procedure without multiplicity adjustment would reject H_0 in (18) if any pairwise test has a P-value less than 0.05. As expected, such a procedure led to inflated type I error rates up to 0.12-0.13. For the pairwise testing procedure adjusted for multiplicity, I used the Bonferroni's correction and adjusted the significance level for each individual test for 3 comparisons. Consequently, a P-value less then 0.0167 would be needed to claim overall significant. The pairwise tests with multiplicity adjustment under RAB design led to actual type I errors rate close to 0.045, which was slightly conservative than the Wald test applied to SMART designs.

Table 5.2. Type I error rates of the Wald test applied to SMART designs and the pairwise tests with and without multiplicity adjustment applied to RAB design at 5% nominal significance (total sample size $n = 400$)

		Pairwise test (RAB)		Wald test (SMART)				
VP	Screen Positive Rate	Unadjusted	Adjusted	S1	S2	S3	S4	S5
VP1	P=0.1	0.125	0.044	0.051	0.047	0.045	0.052	0.052
VP1	P=0.1	0.125	0.044	0.051	0.045	0.046	0.052	0.052
VP1	P=0.1	0.125	0.044	0.051	0.046	0.045	0.052	0.052
VP2	P=0.2	0.122	0.045	0.050	0.049	0.049	0.051	0.051
VP2	P=0.2	0.122	0.045	0.050	0.049	0.049	0.051	0.051
VP2	P=0.2	0.122	0.045	0.050	0.049	0.049	0.051	0.051
VP3	P=0.3	0.120	0.043	0.050	0.051	0.048	0.048	0.052
VP3	P=0.3	0.122	0.045	0.050	0.049	0.049	0.051	0.051
VP3	P=0.3	0.122	0.045	0.050	0.049	0.049	0.051	0.051

Figure 5.5 compares the empirical powers of Wald test applied to SMART designs (S1-S5) versus those of pairwise tests with Bonferroni's correction under RAB design in each outcome scenario. The top panel corresponds to VP1 with screen positive rate equal to 0.1, 0.2 and 0.3, respectively. The empirical powers of Wald test in these settings are consistently greater than those of adjusted pairwise tests in RAB designs. The powers of both Wald test in SMART and adjusted pairwise tests in RAB design increase when the screen positive rate increases (from left to the right). This is because the effect size

increases as the screen positive rate increases given fixed sequence-specific means ϕ_{ijk}'s. S5 achieves the greatest power in each scenario under VP1, which is consistent with the trends of theoretical powers observed in Figure 5.4. The trends of empirical powers under VP2 (second panel) are similar to those under VP1, but the powers under VP2 is lower than those of the same designs under VP1 given a screen positive rate. This is mainly due to smaller effect size under VP2 than that under VP1. Under VP3 (bottom panel), the trends of empirical powers are different from those under VP1 and VP2, as S5 is not always the SMART design with the greatest power. The best SMART design are S5, S1 and S3 for screen positive rate $P = 0.1$, 0.2 and 0.3, respectively, among the 5 designs (S1-S5) we considered in simulation.

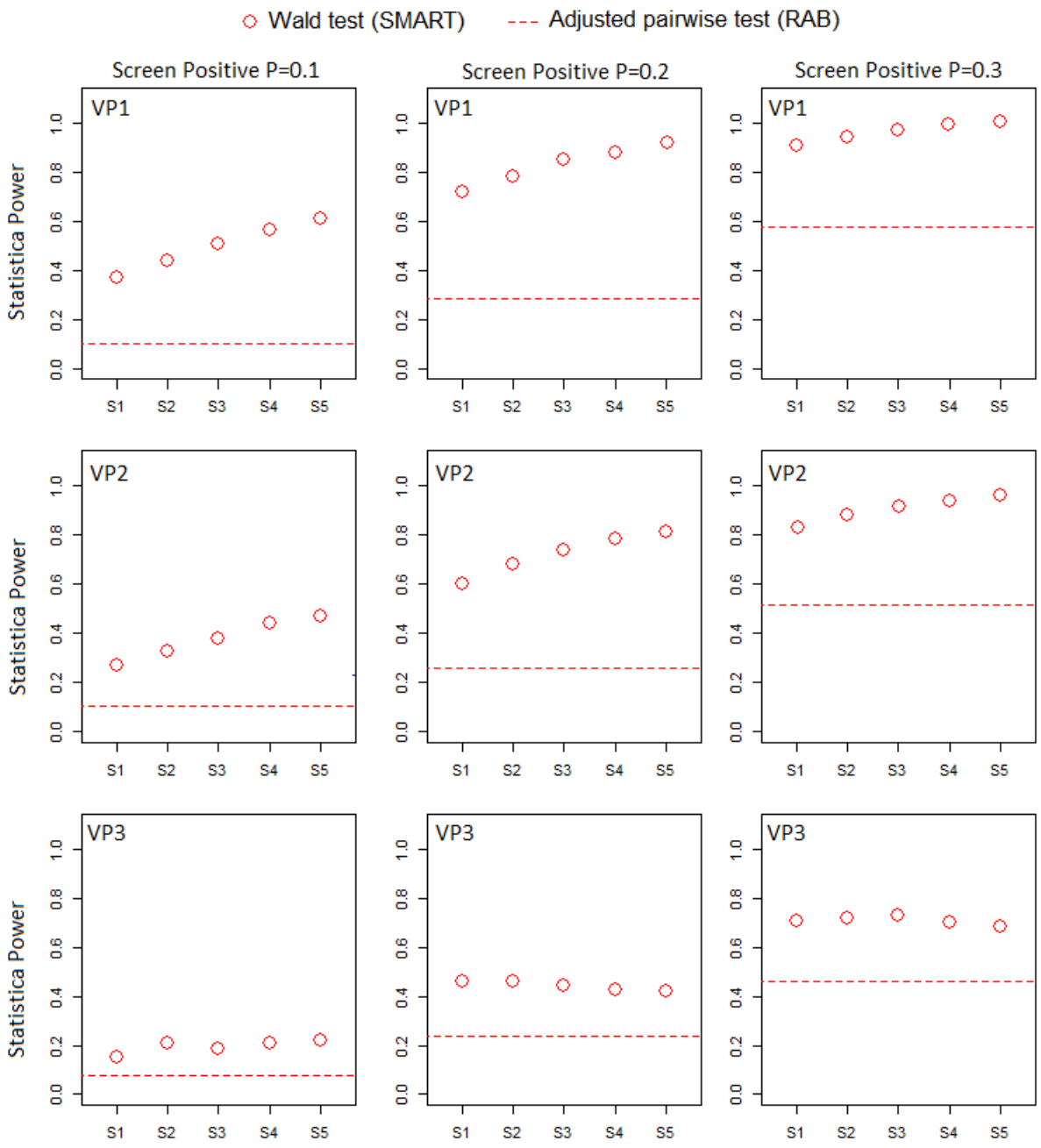


Figure 5.5. Empirical powers of Wald test applied to SMART designs (s1-S5) and adjusted pairwise tests applied to RAB design at 5% nominal significance ($n = 400$).

Figure 5.6. Average primary outcome of patients who participated in SMART and RAB designs considered in the simulation.

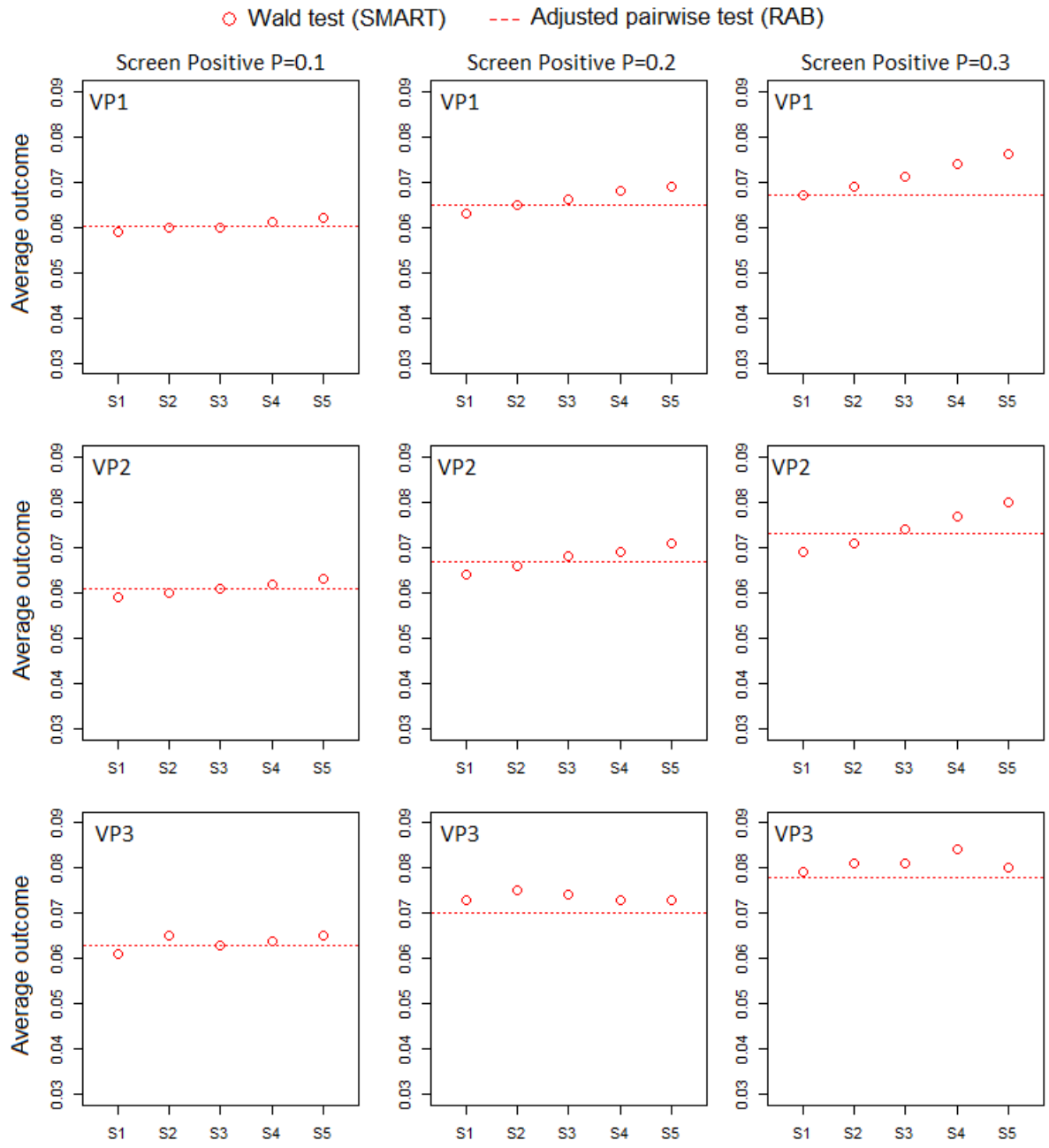


Figure 5.6 gives the average primary outcomes of simulated trials under different SMART designs and the RAB design in 9 outcome scenarios. A RCT design with higher average primary outcome is regarded to be more beneficial for trial participants and thus is more recommendable ethically. To get the average primary outcome of a specific design

in each scenario, I first calculated the average primary outcome over all the subjects in a simulated trial, and then calculated the average outcome across 5000 simulation replicates. Figure 5.6 shows that although some SMART designs (e.g., S1 under VP1 with screen positive rate $P = 0.1$) have lower average primary outcome values comparing to RAB design, I was able to identify at least one SMART design with better average outcome values than the RAB design in all 9 outcome scenarios. In general, the higher proportion of patients referred to treat when the existences of depression are identified by screen, the better average primary outcome a SMART design yields in this case.

4.5 Discussion

SMART is typically considered as a clinical trial design for comparing multi-stage treatment strategies for chronic diseases. For example, in the SMART for comparing two-stage AIs for lymphoma patients (cf. Figure 1.2), a patient received one specific treatment for a period of time in both Stage 1 and Stage 2. In this chapter, we show that SMART can be applied in a more flexible manner. Although the *AHA screen and treatment* strategy and the *screen and notify* strategy only involve one treatment, as long as the strategy involves in multiple clinical decisions and the subsequent decision depends on the prior decision and the results of the prior action, we can applied a SMART design and enjoy the advantage of sequential randomization. Specifically, in the example shown in this chapter, comparing to the traditional RAB design original proposed in protocol, we can design a SMART to answer the same research question with better power and average primary outcome, thus improve the efficiency and quality of patient care.

I explored the possibility of designing a SMART to improve the efficiency and pa-

tient care in clinical trial. I used grad search to identify the randomization probabilities (π_2, π_{222}) of SMART yielded high powers. However, the quantitative relationship between randomization probability and the optimal power given intermediate response rates P_{ij} and sequence-specific parameters of outcome (ϕ_{ijk}, τ_{ijk}) remains unclear. Also, when the design structure becomes complicated, the vector of randomization probabilities used to depict the SMART design will increased from 2 to higher dimension. How to efficiently identify the randomization probabilities that help to achieve the most desired features of a SMART in those situations? These are important problems for my next step study in the future.

Chapter 6 Conclusion and Future Directions

Motivated by developing methods to control the false positive finding in a pre-specified level and improve the power for comparing multiple AIs in SMART, I have proposed methods to address these problems from two different angles, *selection* and *screening*. For selection, I proposed a likelihood-based Wald test that can be applied as a gate-keeping test to select the best AI and study it properties. For screening, I development a method to build simultaneous confidence intervals that can help to identify inferior AI efficiently. As one may view SMART as a clinical trial design at the early phase in a series of experimental studies, selecting the best AI to move it forward and identifying inferior AIs from further investigation represent two ends on this spectrum of experimental series.

The concepts of selection and screening are well-studied in the contexts of clinical trial research for non-adaptive interventions (e.g., Bechhofer, Santner, and Goldsman, 1995; Thall, Simon, and Ellenberg 1988; Cheung 2008). A contribution of my dissertation research is to extend these concepts to the evaluation of adaptive interventions, and thus enhance the practicality of SMART. In the early phase trial where there are potentially many treatment options, one may want to eliminate inferior AIs. I proposed a method to build MCB intervals that can be used to quickly screen SMART data and identify inferior AIs. At the later phase, one may aim to perform a selection trial with a goal to move a single AI to the final confirmatory stage, for which I proposed a gate-keeping Wald test. I also derived a formal sample size calculation formula, which can be used to calculate sample size in designing SMARTs.

I developed and extended the distribution theory of MLE for the value of AIs em-

bedded in general SMART designs. The theory also lends itself the proposed inferential procedures – the gate-keeping test and the MCB simultaneous intervals. In particular, using the asymptotic theory of MLE, we note an important result that the limiting covariance matrix of the MLE is less than full rank. This result is a key element that allows us to establish an efficient gate-keeping test with a null reference distribution with a degrees of freedom $\nu < G - 1$. Without the theory I derived in dissertation, intuitively one might conjecture a null distribution with $G - 1$ degrees of freedom, which would lead to a conservative test. As the design structure of a SMART get more complicated, which usually reflects more intermediate response categories or more stage-specific treatment options, the formula I derived for calculating the exact degree of freedom can help achieve more power in selecting the best AI.

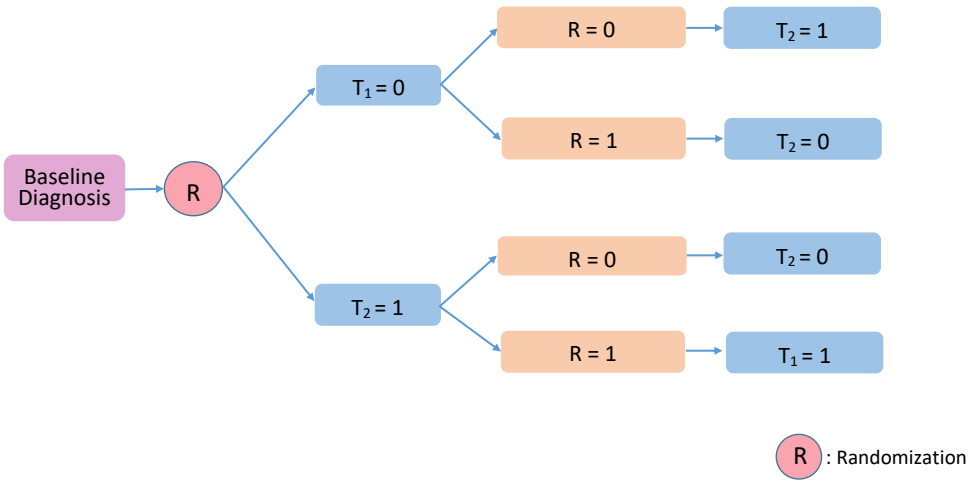


Figure 6.1. An degenerate case of SMART design.

An attractive feature of using the MLE is that the distributional results can be generalized to different types of primary outcomes, such as data with distributions belonging to exponential family. The proof is included in Appendix 1. For the research of chronic diseases, such as cancer, the most commonly used primary endpoint is survival outcome.

In the future, I plan to extend the proposed methods to handle time-to-event data. I have also shown in the dissertation that the proposed methods can be applied to SMARTs with varying design structures and randomization schemes, and the performances were evaluated in finite sample settings using simulations. We note that when the design structure degenerates to an simple case such that only two adaptive interventions were included in the study, as shown in Figure 6.1, the proposed Wald test is equivalent to a single pairwise test described in Chapter 2.3. Another application of the proposed method under degenerate case see the SMART design in Chapter 5. Under the unbalanced randomization scheme, some subgroups of patients may have small randomization probabilities and thus the weights of estimation built based on the inversed randomization probabilities can be big. In such cases, the variation of estimation based on IPWE can be inflated due to the impact of some outlines. While the methods based on MLE is more stable.

A potential limitation of using the MLE is that it requires the full specification of the model, and it may be perceived as restrictive in application. We however note that under normality, the MLE is asymptotically identical to the IPWE, which suggests certain extent of robustness of MLE, at least for continuous outcomes. Furthermore, the proposed gate-keeping procedure is not tied to MLE. Ogabagaber et al. (2016) for example construct a Wald test based on IPWE. As long as we can obtain a consistent estimator for the AI value and the asymptotic variance-covariance of these estimators, we will be able to apply the gate-keeping method. These are certainly topics for further study. Having said that, the results in my dissertation are derived under rather general conditions on the distribution of final primary outcome, with the exponential family being the most prominent example that the theory is applicable to. Thus, the specific procedure studied in my dissertation have applications in very broad settings potentially.

In Chapter 4, I developed an R package to provide an user-friendly statistical software for users to design SMART and analyze SMART data. The package can be used for study design, data analysis and data visualization. So far, I considered SMART as a non-adaptive design, by which the values of design parameters, such as randomization probabilities, are decided at the design stage. Once the trial begins, the values of these design parameters do not change throughout the study. Cheung et al. (2015) proposed a design called Sequential Multiple Assignment Randomization Trial with Adaptive Randomization. Such a design allows clinical trialists to control 3 design parameters based on the results of interim analysis so as to improve the quality of patient care of a SMART. As an extension of dissertation research, I plan to add a function that allows the design parameters to be adapted to the results of interim analysis using the method proposed in Cheung et al (2015) in the near future.

In Chapter 5, I demonstrated the use of SMART to improve the design efficiency in a clinical trial for comparing multiple patient care strategies for depression management. I showed that the randomization probabilities has great impact on the power of SMART. However, the exact quantitative relationship between randomization probabilities and the power of SMART given a design structure and targeted parameters of outcomes remains unclear. In the future, as an extension of the dissertation, I plan to thoroughly examine this relationship under a general SMART design framework, so that the randomization probabilities leading to to optimal power can be identified efficiently.

In my dissertation research, I focused on design and analysis of SMART for comparing multiple AIs, which is challenging given the complicated tree structure of SMART design. The "curse of dimensionality" is a major concern in evaluating AIs embedded in

SMART, as the total number of AIs embedded in SMART increases exponentially as the number of stage, intermediate response categories and stage-specific treatment options increase. The proposed gate-keeping Wald test and the MCB simultaneous intervals are attractive because they are affected by the number of embedded AIs to a lesser extent than pairwise comparison methods with multiplicity adjustments. As a further extension of this research, it will be of interest to develop an adaptive SMART design that can efficiently use the interim analysis results to simplify the tree structure of SMART so as to reduce the dimensionality of SMART data and potentially improve the practicality of SMART design in the world of clinical trial.

Appendix

Appendix 1: Proof of Theorem

Proof of Theorem 1

We assume that $f(y_1|\phi_{ijk}, \tau_{ijk})$ satisfies the regularity conditions asspecified in Theorem 5.39 of van der Varrt (1998). We first prove (3). Noticing that under the standard regularity conditions, we have

$$\sqrt{n}\begin{pmatrix} \hat{\boldsymbol{p}}_i - \boldsymbol{p}_i \\ \hat{\boldsymbol{\phi}}_i - \boldsymbol{\phi}_i \end{pmatrix} \xrightarrow{d} N\left(\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{\boldsymbol{p}_i} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{\boldsymbol{\phi}_i} \end{pmatrix}\right),$$

where $\boldsymbol{\Sigma}_{\boldsymbol{p}_i}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\phi}_i}$ are given in Theorem 1. By the delta-method and using the two equivalent expressions of $\boldsymbol{\theta}_i$ in (1),

$$\begin{aligned} &\sqrt{n}(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i) \\ =&\big(\boldsymbol{A}_i\boldsymbol{\Gamma}_i(\boldsymbol{\phi}_i)|\boldsymbol{A}_i\boldsymbol{\Lambda}(\boldsymbol{p}_i)\big)\sqrt{n}\begin{pmatrix} \hat{\boldsymbol{p}}_i - \boldsymbol{p}_i \\ \hat{\boldsymbol{\phi}}_i - \boldsymbol{\phi}_i \end{pmatrix} + \boldsymbol{O}_p(1) \\ \xrightarrow{d}&N\Big(\mathbf{0}, \boldsymbol{A}_i\big(\boldsymbol{\Gamma}_i(\boldsymbol{\phi}_i)\boldsymbol{\Sigma}_{\boldsymbol{p}_i}\boldsymbol{\Gamma}_i(\boldsymbol{\phi}_i)^T + \boldsymbol{\Lambda}_i(\boldsymbol{p}_i)\boldsymbol{\Sigma}_{\boldsymbol{\phi}}\boldsymbol{\Lambda}_i(\boldsymbol{p}_i)\big)\boldsymbol{A}_i^T\Big) = N\big(\mathbf{0},\ \boldsymbol{\Sigma}_{\boldsymbol{\theta}_i}\big). \end{aligned}$$

Before proving (4), we first establish two lemmas.

Lemma 1: Let $\boldsymbol{A}$ and $\boldsymbol{B}$ be two $k \times k$ real symmetric matrices. Assume $\boldsymbol{A}$ is positive definite and $\boldsymbol{B}$ is positive semi-definite. Then,

$$\text{rank}(\boldsymbol{A} + \boldsymbol{B}) = \text{rank}(\boldsymbol{A}) = k.$$

Proof: For the positive semi-definite matrix $\boldsymbol{A}^{-\frac{1}{2}}\boldsymbol{B}\boldsymbol{A}^{-\frac{1}{2}}$, there exist an orthogonal ma-

trix $\boldsymbol{C}$ such that

$$\boldsymbol{C}\big(\boldsymbol{A}^{-\frac{1}{2}}\boldsymbol{B}\boldsymbol{A}^{-\frac{1}{2}}\big)\boldsymbol{C}^T = \mathrm{diag}\{\lambda_1, \cdots, \lambda_k\},$$

where $\lambda_i \geq 0$, $i = 1, \ldots, k$, are the eigenvalues of $\boldsymbol{A}^{-\frac{1}{2}}\boldsymbol{B}\boldsymbol{A}^{-\frac{1}{2}}$. Therefore,

$$\begin{aligned}\mathrm{rank}(\boldsymbol{A}+\boldsymbol{B}) &= \mathrm{rank}\left(\boldsymbol{I}_k + \boldsymbol{A}^{-\frac{1}{2}}\boldsymbol{B}\boldsymbol{A}^{-\frac{1}{2}}\right)\\ &= \mathrm{rank}\left(\boldsymbol{I}_k + \boldsymbol{C}\boldsymbol{A}^{-\frac{1}{2}}\boldsymbol{B}\boldsymbol{A}^{-\frac{1}{2}}\boldsymbol{C}^T\right)\\ &= \mathrm{rank}\left(\mathrm{diag}\{1+\lambda_1, \ldots, 1+\lambda_k\}\right) = k.\end{aligned}$$

Lemma 2: Let $\boldsymbol{A}_i$ be defined as in equation (2). Then,

$$\mathrm{rank}(\boldsymbol{A}_i) = \sum_{j=1}^{J_i} K_{ij} - J_i + 1 = m_i - J_i + 1.$$

Proof: We apply the principle of mathematical induction to J_i. For such purpose, we write $\boldsymbol{A}_i$ as $\boldsymbol{A}_i(K_{i1}, \ldots, K_{iJ_i})$. If $J_i = 2$, then

$$\boldsymbol{A}_i(K_{i1}, K_{i2}) = \big(\boldsymbol{I}_{K_{i1}} \otimes \boldsymbol{1}_{K_{i1}} | \boldsymbol{1}_{K_{i1}} \otimes \boldsymbol{I}_{K_{i2}}\big),$$

which, after some elementary operations for block matrices, becomes

$$\begin{pmatrix} \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{I}_{K_{i2}} \\ \boldsymbol{0} & \boldsymbol{I}_{K_{i1}-1} \otimes \boldsymbol{1}_{K_{i2}} & \boldsymbol{0} \end{pmatrix}.$$

Thus

$$\mathrm{rank}\left(\boldsymbol{A}_i(K_{i1}, K_{i2})\right) = \mathrm{rank}\left(\boldsymbol{I}_{K_{i2}}\right) + \mathrm{rank}\left(\boldsymbol{I}_{K_{i1}-1} \otimes \boldsymbol{1}_{K_{i2}}\right) = K_{i2} + K_{i1} - 1.$$

Suppose the conclusion holds for J_i, consider now the case of $J_i + 1$. Denote

$$K_i' = \sum_{j=2}^{J_i+1} K_{ij} \ .$$

Since after some elementary operations for block matrices

$$\boldsymbol{A}_i\big(K_{i1}, \ldots, K_{iJ_i}, K_{i,J_i+1}\big) = \big(\boldsymbol{I}_{K_{i1}} \otimes \mathbf{1}_{K_i'} | \mathbf{1}_{K_{i1}} \otimes \boldsymbol{A}_i(K_{i1}, \ldots, K_{i,J_i+1})\big)$$

becomes

$$\begin{pmatrix} \mathbf{0} & \mathbf{0} & \boldsymbol{A}_i(K_{i2}, \ldots, K_{i,J_i+1}) \\ \mathbf{0} & \boldsymbol{I}_{K_{i1}-1} \otimes \mathbf{1}_{K_i'} & \mathbf{0} \end{pmatrix},$$

we have

$$\begin{aligned} \operatorname{rank}\big(\boldsymbol{A}_i(K_{i1}, \ldots, K_{i,J_i+1})\big) &= \operatorname{rank}\big(\boldsymbol{A}_i(K_{i2}, \ldots, K_{i,J_i+1})\big) + \operatorname{rank}\big(I_{K_{i1}-1} \otimes \mathbf{1}_{K_i'}\big) \\ &= K_i' - J_i + 1 + K_{i1} - 1 \\ &= \sum_{j=1}^{J_i+1} K_{ij} - (J_i + 1) + 1, \end{aligned}$$

which proves the claim for $J_i + 1$.

We now prove (4). By Lemma 1,

$$\operatorname{rank}\big(\boldsymbol{\Lambda}_i(\boldsymbol{p}_i)\boldsymbol{\Sigma}_{\boldsymbol{\phi}_i}\boldsymbol{\Lambda}_i(\boldsymbol{p}_i) + \boldsymbol{\Gamma}_i(\boldsymbol{\phi}_i)\boldsymbol{\Sigma}_{\boldsymbol{p}_i}\boldsymbol{\Gamma}_i(\boldsymbol{\phi}_i)\big) = m_i,$$

hence of full rank. Then, by Lemma 2,

$$\begin{aligned} \operatorname{rank}\big(\boldsymbol{\Sigma}_{\boldsymbol{\theta}_i}\big) &= \operatorname{rank}\big(\boldsymbol{A}_i\big(\boldsymbol{\Lambda}_i(\boldsymbol{p}_i)\boldsymbol{\Sigma}_{\boldsymbol{\phi}_i}\boldsymbol{\Lambda}_i(\boldsymbol{p}_i) + \boldsymbol{\Gamma}_i(\boldsymbol{\phi}_i)\boldsymbol{\Sigma}_{\boldsymbol{p}_i}\boldsymbol{\Gamma}_i(\boldsymbol{\phi}_i)\big)\boldsymbol{A}_i^T\big) \\ &= \operatorname{rank}(\boldsymbol{A}_i) = \sum_{j=1}^{J_i} K_{ij} - J_i + 1. \end{aligned}$$

Proof of Theorem 2

By (3) in Theorem 1, under H_0, $Q \xrightarrow{d} \chi^2_\nu$. By a contiguity argument, under the local alternatives $\{\boldsymbol{\Theta}_n\}$ which satisfies (8), $Q \xrightarrow{d} \chi^2_\nu(\lambda^*)$. We now verify that the degrees of freedom formula (7). Let $G = \sum_{i=1}^I G_i$ and $m = \sum_{i=1}^I m_i$. Define an $G \times m$ matrix $\boldsymbol{A}$ as

$$\boldsymbol{A} = \text{bdiag}\{\boldsymbol{A}_i; i = 1, \ldots, I\}.$$

Without loss of generality, consider an $(G-1) \times G$ contrast matrix

$$\boldsymbol{C} = \big(\mathbf{1}_{G-1} | - I_{G-1}\big).$$

By subtracting the first row from the remaining $(G-1)$ rows in $\boldsymbol{A}$, and then subtracting the first column from the remaining columns (all of these are elementary operations), $\boldsymbol{A}$ is converted to

$$\begin{pmatrix} \mathbf{1} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{B} \end{pmatrix},$$

and check that $(\mathbf{0} \mid \boldsymbol{B}) = \boldsymbol{CA}$ holds. Then,

$$\text{rank}(\boldsymbol{A}) = 1 + \text{rank}(\boldsymbol{B}) = 1 + \text{rank}(\boldsymbol{CA}).$$

Therefore, the degrees of freedom of χ^2_ν test is

$$\nu = \text{rank}\big(\boldsymbol{CA}\big) = \text{rank}(\boldsymbol{A}) - 1 = \sum_{i=1}^{I}\sum_{j=1}^{J_i} K_{ij} - \sum_{i=1}^{I} J_i + I - 1.$$

Proof of Theorem 3

Let $g^* = \underset{1\le i\le G}{\operatorname{argmax}}\,\theta_i$, where ties can be broken in any fashion without affecting the validity of the proof. Consider

$$E = \left\{\hat{\theta}_i - \theta_i - \delta_{g^*}\hat{\sigma}_{ig^*}/\sqrt{n} \le \hat{\theta}_{g^*} - \theta_{g^*} \le \hat{\theta}_i - \theta_i + \delta_{g^*}\hat{\sigma}_{ig^*}/\sqrt{n};\ i \ne g^*\right\}.$$

By the construction of δ_{g^*},

$$\lim_{n\to\infty} P(E) = 1 - \alpha.$$

Since

$$\hat{\theta}_{g^*} - \theta_{g^*} \ge \hat{\theta}_i - \theta_i - \delta_{g^*}\hat{\sigma}_{ig^*}/\sqrt{n}$$

for all $i \ne g^*$ is equivalent to

$$\hat{\theta}_{g^*} - \hat{\theta}_i + \delta_{g^*}\hat{\sigma}_{ig^*}/\sqrt{n} \ge \theta_{g^*} - \theta_i \ge 0$$

for all $i \ne g^*$, we conclude that $g^* \in \mathcal{B}$.

From

$$\hat{\theta}_{g^*} - \theta_{g^*} \le \hat{\theta}_i - \theta_i + \delta_{g^*}\hat{\sigma}_{ig^*}/\sqrt{n},$$

we have

$$\theta_i - \theta_{g^*} \le \hat{\theta}_i - \hat{\theta}_{g^*} + \delta_{g^*}\hat{\sigma}_{ig^*}/\sqrt{n},$$

and noticing that $\theta_i - \theta_{g^*} \le 0$, we conclude that

$$\begin{aligned} E &\subseteq \left\{g^* \in \mathcal{B}, \theta_i - \theta_{g^*} \le \min\left\{0, \hat{\theta}_i - \hat{\theta}_{g^*} + \delta_{g^*}\hat{\sigma}_{ig^*}/\sqrt{n}\right\}, i \ne g^*\right\} \\ &\subseteq \left\{\theta_i - \theta_{g^*} \le U_i, i = 1, \ldots, G\right\}. \end{aligned}$$

Similarly,

$$\begin{aligned} E &\subseteq \Big\{ g^* \in \mathcal{B}, \theta_i - \theta_{g^*} \geq \hat{\theta}_i - \hat{\theta}_{g^*} - \delta_{g^*} \hat{\sigma}_{ig^*}/\sqrt{n}, i \neq g^* \Big\} \\ &\subseteq \Big\{ \theta_i - \theta_{g^*} \geq L_i, i = 1, \ldots, G \Big\}. \end{aligned}$$

Thus,

$$E \subseteq \Big\{ L_i \leq \theta_i - \theta_{g^*} \leq U_i, i = 1, \ldots, G \Big\},$$

and

$$P\Big(L_i \leq \theta_i - \max_{1 \leq j \leq G} \theta_j \leq U_i, i = 1, \ldots, G \Big) = P\Big(L_i \leq \theta_i - \theta_{g^*} \leq U_i, i = 1, \ldots, G \Big) \geq P(E).$$

Therefore,

$$\liminf_{n \to \infty} P\Big(L_i \leq \theta_i - \max_{1 \leq j \leq G} \theta_j \leq U_i, i = 1, \ldots, G \Big) \geq \lim_{n \to \infty} P(E) = 1 - \alpha.$$

Appendix 2. Specification of $\boldsymbol{\phi}_{ijk}$'s in simulation

We provide an example of generating the sequence-specific mean outcome ϕ_{ijk} in the simulations under design structure 1 (DS1) and balanced randomization scheme (BR) with value pattern 1 (VP1). There are 8 possible treatment sequences in this setting and the sequence-specific means ϕ_{ijk} can be expressed as a set of linear functions of $\boldsymbol{\beta}$ as follows,

$$\begin{aligned}
\phi_{(111)} &= \beta_0 \\
\phi_{(112)} &= \beta_0 + \beta_3 \\
\phi_{(121)} &= \beta_0 + \beta_2 \\
\phi_{(122)} &= \beta_0 + \beta_2 + \beta_3 + \beta_6 \\
\phi_{(211)} &= \beta_0 + \beta_1 \\
\phi_{(212)} &= \beta_0 + \beta_1 + \beta_3 + \beta_5 \\
\phi_{(221)} &= \beta_0 + \beta_1 + \beta_2 + \beta_4 \\
\phi_{(222)} &= \beta_0 + \beta_1 + \beta_2 + \beta_3 + \beta_4 + \beta_5 + \beta_6 + \beta_7.
\end{aligned}$$

Also, the value of an AI in this setting can be calculated by

$$\theta_{i;k_{i1},k_{i2}} = P_{i1}\phi_{i1k_{i1}} + P_{i2}\phi_{i2k_{i2}},$$

where $k_{i1} \in (0,1)$ and $k_{i2} \in (0,1)$ for $i = 1,2$. Thus, the targeted AI values can be

expressed as

$$\boldsymbol{\Theta}^*_{\boldsymbol{\beta}} = \begin{pmatrix} \theta_{1;1,1} \\ \theta_{1;1,2} \\ \theta_{1;2,1} \\ \theta_{1;2,2} \\ \theta_{2;1,1} \\ \theta_{2;1,2} \\ \theta_{2;2,1} \\ \theta_{2;2,2} \end{pmatrix} = \begin{pmatrix} \beta_0 + \frac{1}{3}\beta_2 \\ \beta_0 + \frac{1}{3}\beta_2 + \frac{1}{3}\beta_3 + \frac{1}{3}\beta_6 \\ \beta_0 + \frac{1}{3}\beta_2 + \frac{2}{3}\beta_3 \\ \beta_0 + \frac{1}{3}\beta_2 + \beta_3 + \frac{1}{3}\beta_6 \\ \beta_0 + \beta_1 + \frac{1}{3}\beta_2 + \frac{1}{3}\beta_4 \\ \beta_0 + \beta_1 + \frac{1}{3}\beta_2 + \frac{1}{3}\beta_3 + \frac{1}{3}\beta_4 + \frac{1}{3}\beta_5 + \frac{1}{3}\beta_6 + \frac{1}{3}\beta_7 \\ \beta_0 + \beta_1 + \frac{1}{3}\beta_2 + \frac{2}{3}\beta_3 + \frac{1}{3}\beta_4 + \frac{2}{3}\beta_5 \\ \beta_0 + \beta_1 + \frac{1}{3}\beta_2 + \beta_3 + \frac{1}{3}\beta_4 + \beta_5 + \frac{1}{3}\beta_6 + \frac{1}{3}\beta_7 \end{pmatrix}. \tag{22}$$

We added $\boldsymbol{\beta}$ on the right bottom of $\boldsymbol{\Theta}^*$ in (22) to indicate that the value of $\boldsymbol{\Theta}^*$ only depends on the values of $\boldsymbol{\beta}$. With VP1, we have

$$\theta_{1;1,1} = \theta_{1;1,2} = \theta_{1;2,1} = \theta_{1;2,2} < \theta_{2;1,1} = \theta_{2;1,2} = \theta_{2;2,1} = \theta_{2;2,2}. \tag{23}$$

From (22) and (23) we know that any set of $\boldsymbol{\beta}$ satisfying $\beta_1 > 0$ and $\beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = 0$ can be used to build a SMART under DS1 and with VP1. Similarly, we can go further to restrict the values of $\boldsymbol{\beta}$ under BR and $\Delta = 0.05$. First, we build the target equation. The covariance between two AI values in this setting can be calculated by

$$\begin{aligned} cov\left(\theta_{i;k_{i1},k_{i2}}, \theta_{i';k_{i'1},k_{i'2}}\right) = & \frac{P_{i1}P_{i2}(\phi_{ik_{i2}} - \phi_{ik_{i1}})(\phi_{i'k_{i'2}} - \phi_{i'k_{i'1}})I\{T_i = T_{i'}\}}{\pi_i} \\ & + \frac{P_{i1}\sigma^2 I\{S_{i1k_{i1}} = S_{i'1k_{i'1}}\}}{\pi_i \pi_{i1k_{i1}}} \\ & + \frac{P_{i2}\sigma^2 I\{S_{i2k_{i2}} = S_{i'2k_{i'2}}\}}{\pi_i \pi_{i2k_{i2}}}, \end{aligned} \tag{24}$$

where $i, i', k_{i1}, k_{i2}, k_{i'1}, k_{i'2} = 1, 2$; $I\{E\} = 1$ when event E occurs and $I\{E\} = 0$ otherwise. The value of $I\{.\}$ in (24) changes according to the relationship between two AIs of $d_{i;k_{i1},k_{i2}}$ and $d_{i';k_{i'1},k_{i'2}}$. For example, when two AIs are complete overlapped, we have $I\{T_i = T_{i'}\} = I\{S_{i1k_{i1}} = S_{i'1k_{i'1}}\} = I\{S_{i2k_{i2}} = S_{i'2k_{i'2}}\} = 1$ so that (24) is the variance of an AI. Meanwhile, when both AIs suggest the same stage-1 treatment, but different

treatments for either responders or non-responders at stage 2, we have $I\{T_i = T_{i'}\} = 1$ and $I\{S_{i1k_{i1}} = S_{i'1k_{i'1}}\} = I\{S_{i2k_{i2}} = S_{i'2k_{i'2}}\} = 0$. In this way, we can calculate all the elements of target $\boldsymbol{\Sigma}^*$ in (9) by 8×8 known functions of $\{\pi_i, \pi_{ijk}, P_{i1}, P_{i2}, \phi_{ijk}, \sigma_{ijk}\}$. In this case, we have $\pi_i = \pi_{ijk} = 0.5$, $(P_{i1}, P_{i2}) = (\frac{2}{3}, \frac{1}{3})$ and $\sigma_{ijk} = 10$, for $i, j, k = 1, 2$, so the value of $\boldsymbol{\Sigma}^*$ only depends on $\boldsymbol{\beta}$. With $\Delta = 0.05$, we can build the target equation as

$$(\boldsymbol{C}\boldsymbol{\Theta}^*_{\boldsymbol{\beta}})^T \big(\boldsymbol{C}\boldsymbol{\Sigma}^*_{\boldsymbol{\beta}}\boldsymbol{C}^T\big)^{-}(\boldsymbol{C}\boldsymbol{\Theta}^*_{\boldsymbol{\beta}}) = 0.05, \tag{25}$$

where the contras matrix

$$\boldsymbol{C} = \begin{pmatrix} 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & -1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & -1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & -1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & -1 \end{pmatrix}.$$

By solving (25), we obtained a set of $\boldsymbol{\beta} = (0, 4.47, 0, 0, 0, 0, 0, 0)$, which were used to simulate the SMART data under DS1 and BR, with VP1 and $\Delta = 0.05$. In each simulated SMART data based on $\boldsymbol{\beta} = (0, 4.47, 0, 0, 0, 0, 0, 0)$, there were 8 possible sequences and the sequences-specific means were $\Phi = (0, 0, 0, 0, 4.7, 4.7, 4.7, 4.7)$. The solution of target equation (25) is not unique, since β_0 can take any value of $(-\infty, +\infty)$. However with the goal of simulating a SMART with desired features, any set of $\boldsymbol{\beta}$ satisfying (25) can be selected to use. We adopted $\beta_0 = 0$ in the simulation. The values of $\boldsymbol{\beta}$'s for all the scenarios in the simulations are provided in the table 2.3.

References

[1] American Society of Clinical Oncology: Update of recommendations for the use of hematopoietic colony-stimulating factors: Evidence-based clinical practice guidelines. *Journal Clinical Oncology* 14, 1957-1960, 1996.

[2] Bechhofer RE, Santner TJ, Goldsman DM. Design and analysis of experiments for statistical selection, screening, and multiple comparison. 1995; New York: Wiley.

[3] Brown B. (1980). "The crossover experiment for clinical trials," *Biometric*, 36, 69-79.

[4] Bush DE, Ziegelstein RC, Patel UV, et al. Post-myocardial Infarction Depression Summary: Agency for Healthcare Research and Quality;2005. 123.

[5] Carney RM, Freedland KE. Depression in patients with coronary heart disease. Am. J. Med. Nov 2008; 121(11 Suppl 2):S20-27.

[6] Chakraborty B, Collins L, Strecher V and Murphy S. (2009), "Developing multicomponent interventions using fractional factorial designs," *Statistics in Medicine*, 28, 2687-2708.

[7] Chakraborty B. and Murphy. (2014), "Dynamic Treatment Regimes," *Annal of Applied Statistics*, XX, XXX-XXX.

[8] Chakraborty, B., Strecher, V., and Murphy, S. (2010), "Inference for nonregular parameters in optimal dynamic treatment regimes," *Statistical Methods in Medical Research*, 19, 317–343.

[9] Cheung YK. Simple sequential boundaries for treatment selection in multi-armed randomized clinical trials with a control. Biometrics. 2008; 64:940–949.

[10] Cheung YK. Sequential implementation of stepwise procedures for identifying the maximum tolerated dose. Journal of the American Statistical Association. 2007; 102:1448–1461.

[11] Cheung, Y. K., Chakraborty, B., and Davidson, K. W. (2015), "Sequential multiple assignment randomized trials (SMART) with adaptive randomization for quality improvement in depression treatment program," *Biometrics*, 71, 450-459.

[12] Coffey C, Levin B Clark C, Timmerman C, Wittes J et al. (2012), "Overiew, hurdles. And future work in adaptive designs: perspectives from an NIH-funded workshop," *Clinical Trials*, 9, 671-680.

[13] Collins L, Murphy S and Bierman K. (2004), "A conceptual framework for adaptive preventive interventions," *Prevention Science*, 5, 185-196.

[14] Czuczman MS, Grillo-Lopez AJ, White CA, et. al. (1999), "Treatment of patients with low-grade B-cell lymphoma with the combination of chimeric anti-CD20 monoclonal antibody and CHOP chemotherapy." *Journal of Clinical Oncology*, 17, 268-276.

[15] Edwards, D. G., and Hsu, J. C. (1983), "Multiple comparison with the best treatment," *Journal of the American Statistical Association*, 78, 965-971.

[16] Fisher R. (1926), "The arrangement of field experiments," *Journal of the Ministry of Agriculture of Great Britain*, 33, 503-513.

[17] Gallo P, Chuang-Stein C, Dragalin V, Gaydos B, Krams M, Pinheir J; PhRMA working group. (2006), "Adaptive designs in clinical drug development-an Executive Summary of the PhRMA Working Group," *Journal of Biopharmaceutical Statistics*, 16, 275-283.

[18] Garrett-Mayer E. (2006), "The continual reassessment method for dose-finding studies: a tutorial," *Clinical Trials*, 3, 57-71.

[19] Giachetti R, Marcelli V, Cifuentes K and Rojas. (2013), "An agent-based simulation model of human-robot team performance in military environments," *Systems Engineering*, 16, 15-28.

[20] Graham I, Atar D, Borch-Johnsen K, et al. European guidelines on cardiovascular disease prevention in clinical practice: executive summary. Eur. Heart J. Oct 2007;28(19):2375-2414.

[21] Guyatt GH, Haynes RB, Jaeschke RZ, et al. Users' Guides to the Medical Literature: XXV. Evidencebased medicine: principles for applying the Users' Guides to patient care. Evidence-Based Medicine Working Group. JAMA. Sep 13 2000;284(10):1290-1296.

[22] Habermann T. M. Weller E. A., Morrison V. A., Gascoyne R. D., Cassileth P. A. Cohn, J. B., Dakhil S. R., Woda B., Fisher R. I., Peterson B. A., Horning S. J. (2006), "Rituximab-CHOP versus CHOP along with maintenance Rituximab in older patients with diffuse large B-Cell Lymphoma," *Journal of Clinical Oncology*, 24, 3121-3127.

[23] Hochberg Y. and Tamhane A.C. (1987), "Multiple Comparison Procedures," *New York, Wiley.*

[24] Holland C and Cravens D. (1973), "Fractional factorial experimental designs in marketing Rresearch," *Journal of Marketing Research*, 10, 270-276.

[25] Holloway S.T., Laber E.B., Linn K.A., Zhang B., Davidian M., and Tsiatis A.A. (2017), "DynTxRegime: A comprehensive package for analysis of dynamic treatment regimes (R)," *Innovative Methods Program for Advancing Clinical Trials (IMPACT)*, v3.1 ed.

[26] Hsu, J. C. (1981), "Simultaneous confidence intervals for all distances from the best," *The Annals of Statistics*, 9, 1026-1034.

[27] Hsu, J. C. (1984), "Constrained simultaneous confidence intervals for multiple comparison with the best," *The Annals of Statistics*, 12, 1136-1144.

[28] Kosorok and Moodie. (1988), "DTR book," *Biometrika*, 75, 303-310.

[29] Lanceford J, Davidian M and Tsiatis A. (2002), "Estimation of survival distributions of treatment policies in two-stage randomization designs in clinical trials," *Biometrics*, 58, 48-57.

[30] Lavori, P. W., and Dawson, R. (2007), "Improving the efficiency of estimation in randomized trials of adaptive treatment strategies," *Clinical Trials*, 4, 297-308.

[31] Lichtman JH, Bigger JT, Jr., Blumenthal JA, et al. Depression and coronary heart disease: recommendations for screening, referral, and treatment: a science advisory from the American Heart Association Prevention Committee of the Council on Cardiovascular Nursing, Council on Clinical Cardiology, Council on Epidemiology and Prevention, and Interdisciplinary Council on Quality of Care and Outcomes Research: endorsed by the American Psychiatric Association. Circulation. Oct 21 2008;118(17):1768-1775.

[32] Linn, K.A., Laber, E.B. and Stefanski, L.A. (2015), "iqLearn: Interactive Q-Learning in R," *Journal of Statistical Software*, 64, 1–25.

[33] Maca J, Bhattacharya S, Dragalin V. Gallo P, Krams M. Adaptive Seamless Phase II/III designs-background, operational aspects and examples. Information Journal. 40: 463-473.

[34] Meier U. A note on the power of Fisher's least significant difference procedure. Pharmaceutical Statistics. 2006; 5:253–263.

[35] Murphy, S. A. (2005a)," An experimental design for the development of adaptive treatment strategies," *Statistics in Medicine*, 24, 1455-1481.

[36] Murphy, S. A., van der Laan, M. J., Robin, J. M., and CPPRG. (2001), "Marginal mean models for dynamic regimes," *Journal of American Statistical Association*, 96, 1410-1423.

[37] Nahum-Shani, I., Qian, M., Almirall, D., Pelham, W. E., Gnagy, B., Fabiano, G., Waxmonsky, J., Yu, J., and Murphy, S. (2012), "Experimental design and primary data analysis for developing adaptive interventions," *Psychological Methods*, 17, 457-477.

[38] National Institute for H, Clinical E. Depression in Adults with Chronic Physical Health Problems. London: National Institute for Health and Clinical Excellence;2009.

[39] Nicholson A, Kuper H, Hemingway H. Depression as an aetiologic and prognostic factor in coronary heart disease: a meta-analysis of 6362 events among 146 538 participants in 54 observational studies. Eur. Heart J. Dec 2006;27(23):2763-2774.

[40] Oetting A.I., Levy J.A., Weiss R.D. and Murphy S.A. (2011), "Statistical methodology for a SMART design in the development of adaptive treatment strategies," In Causality and Psychopathology: Finding the Determinants of Disorders and their Cures American Psychiatric Publishing: Arlington, VA, 179–205.

[41] Ogbagaber, S. B., Karp, J., and Wahed, A. S. (2016), "Design of sequentially randomized trials for testing adaptive treatment strategies," *Statistics in Medicine*, 35, 840-858.

[42] Orellana, L., Rotnitzky, A., and Robins, J. M. (2010), "Dynamic regime marginal structural mean models for estimation of optimal dynamic treatment regimes, Part I: main content," *International Journal of Biostatistics*, 6, Article.

[43] Post-Myocardial Infarction Depression Clinical Practice Guideline P. AAFP Guideline for the Detection and Management of Post-Myocardial Infarction Depression. Ann Fam Med. January 1, 2009 2009;7(1):71-79.

[44] Proschan MA. (2009), "Sample size re-estimation in clinical trials," *Biometrical Journal*, 51, 348-357.

[45] Robins, J. M. (1986), "A new approach to causal inference in mortality studies with sustained exposure periods-application to control to the health worker survivor effect," *Computers and Mathematics with Applications*, 14, 1393-1512.

[46] Rush AJ, Fava M, Wisniewski SR, Lavori PW, Trivedi MH, Sackeim HA, Thase ME, Nierenberg AA, Quitkin FM, Kashner TM, Kupfer DJ, Rosenbaum JF, Alpert J, Stewart JW, McGrath PJ, Biggs MM, Shores-Wilson K, Lebowitz BD, Ritz L, Niederehe G,

STAR*D Investigators Group. Sequenced treatment alternatives to relieve depression (STAR*D): rationale and design. Controlled Clinical Trial. 2004; 25:119-142.

[47] Stafford L, Berk M, Reddy P, Jackson HJ. Comorbid depression and health-related quality of life in patients with coronary artery disease. J. Psychosom. Res. 2007;62(4):401-410.

[48] Strecher, V. J., McClure, J. B., Alexander, G. L., Chakraborty, B., Nair, V. N., Greene, S. M., Collins, L. M., Carlier, C. C., Wiese, C. J., Little, R. J., Pomerleau, C. S., and Pomerleau, O. F. (2008), "Web-based smoking-cessation programs: results of a randomized trial," *American Journal or Preventive Medicine*, 34, 373-381.

[49] Tang, X. and Melguizo, M (2015), "DTR: an R package for estimation and comparison of survival outcomes of dynamic treatment regimes," *Journal of Statistical Software*, 65, 1–28.

[50] Thall PF, Simon R, Ellenberg SS. Two-stage selection and testing designs for comparative clinical trials. Biometrika. 1988; 75:303–310.

[51] Thall, PF, Wooten LH, Logothetis CJ, Millikan RE, Tannir NM. Bayesian and frequentist two-stage treatment strategies based on sequential failure times subject to interval censoring. Statistics in Medicine. 2007; 26:4687–4702.

[52] Van der Laan MJ and Petersen ML. (2007), "Statistical learning of origin-specific statistically optimal individualized treatment rules," 3: article 6.

[53] Van der Varrt, A.W. (1998), "Asymptotic Statistics," New York, Cambridge University Press.

[54] Von Korff M, Ormel J, Katon W, Lin EH. Disability and depression among high utilizers of health care. A longitudinal analysis. Arch. Gen. Psychiatry. Feb 1992;49(2):91-100.

[55] Wahed A and Tsiatis A. (2004), "Optimal estimator for the survival distribution and related quantities for treatment policies in two-stage randomized designs in clinical trials," *Biometrics*, 60, 124-133.

[56] Wallace, P., Moodie, M. and Stephens, A., (2017). Dynamic treatment regimen estimation via regression-based technique: Introducing R package DTRreg. *Journal of Statistical Software*, 80, doi: 10.18637/jss.v080.i02.

[57] Wang SJ, Hung HM and O'Neill RT. 2009. Adaptive patient enrichment design in therapeutic trials. Biometrical Journal. 51, 358-374.

[58] Zhang L and Rosenberger WF. (2012), "Adaptive randomization in clinical trials," *Design and Analysis of Experiments, Special Designs and Applications*, 3, 251-282.